Bridging Trust: Combating Fakes and Frauds with Robust Graph Learning

Keynote Speech @ PAKDD – RAFDA workshop 2024

Cheng-Te Li

Dept. of CSIE

National Cheng Kung University

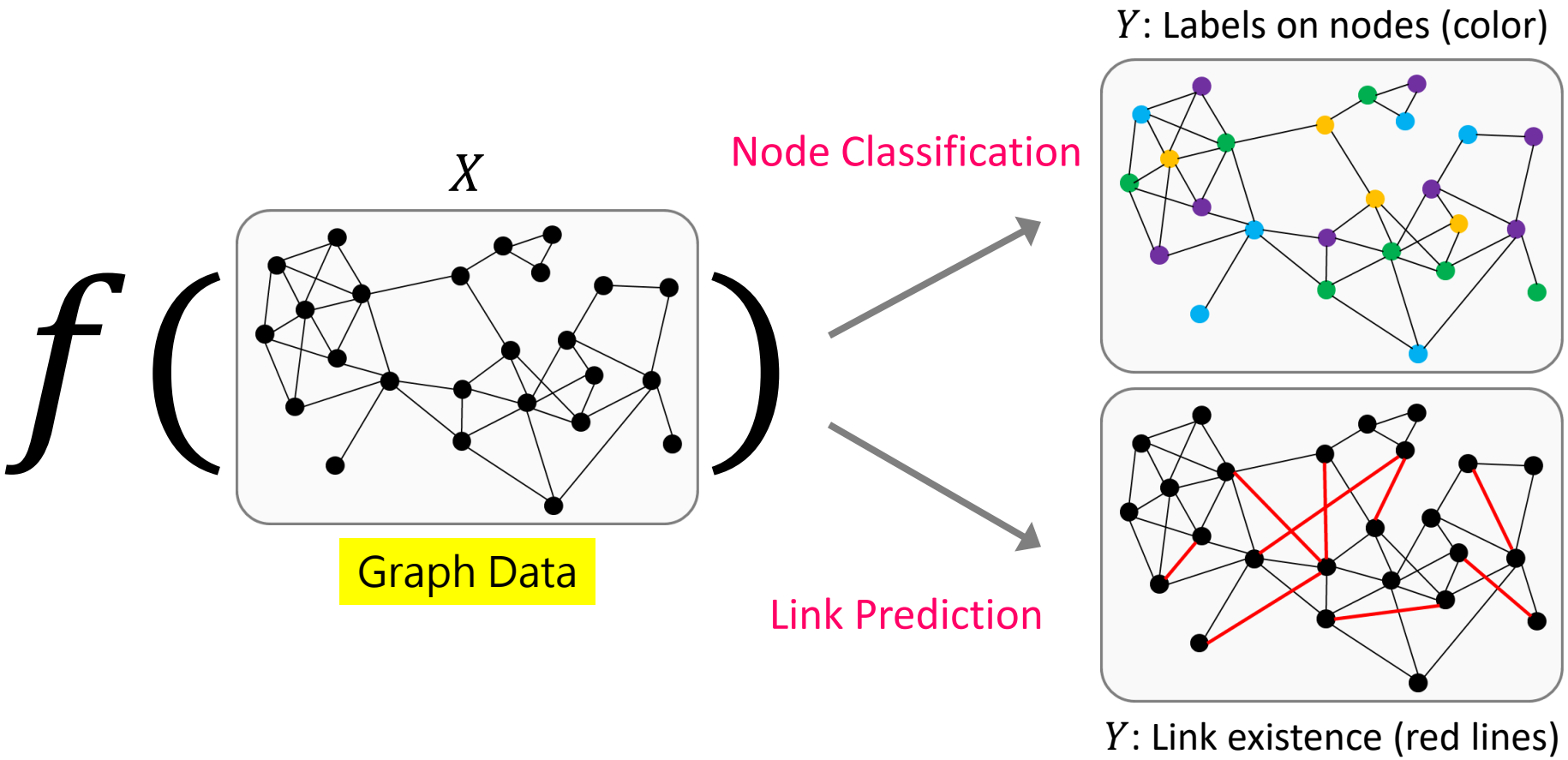
chengte@ncku.edu.tw

Fakes and Frauds

- Both involve **deception** or **misrepresentation**
  - Deliberately deceive the recipient into believing something that is not true
- **Fakes**
  - Counterfeit items or identities without the permission of the original creator
  - E.g., fake news
- **Frauds**
  - A broader range of deceptive practices aimed at financial or personal gain
  - E.g., customs frauds, illicit food factories, link frauds

Graph Machine Learning (GraphML)

$$f(X) \rightarrow Y$$



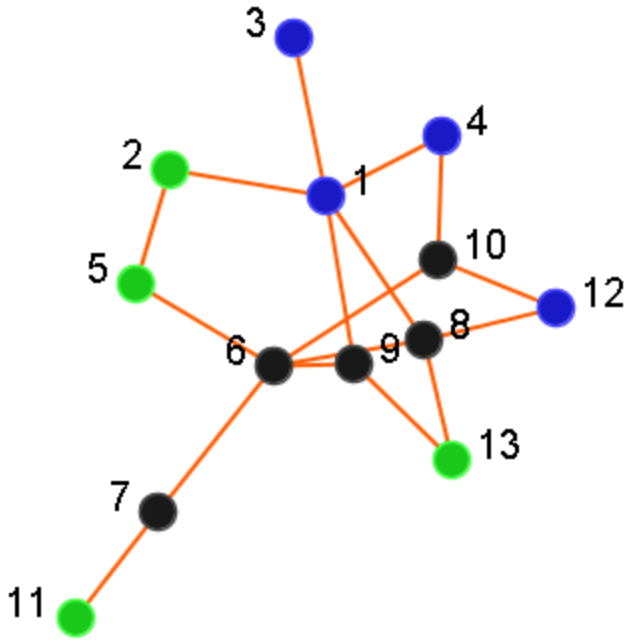
Node Classification in GraphML

COVID-19 Test

Public Opinion

Online Advertising

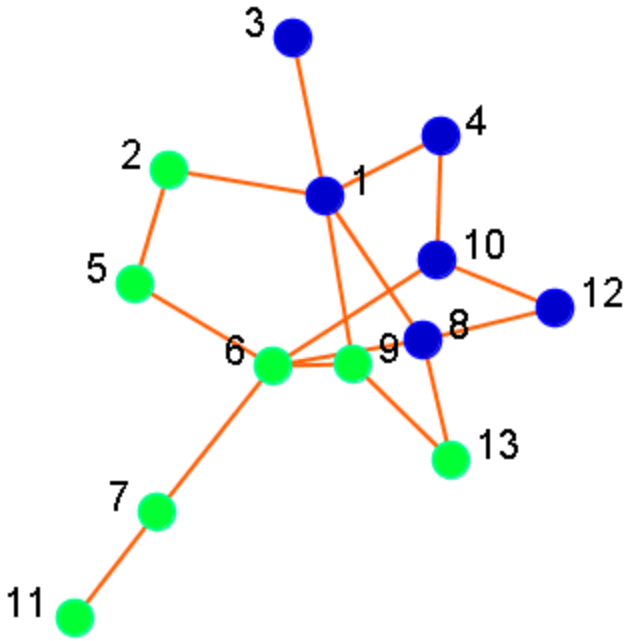
●: Gaming
●: Cosmetics
●: ? Unknown



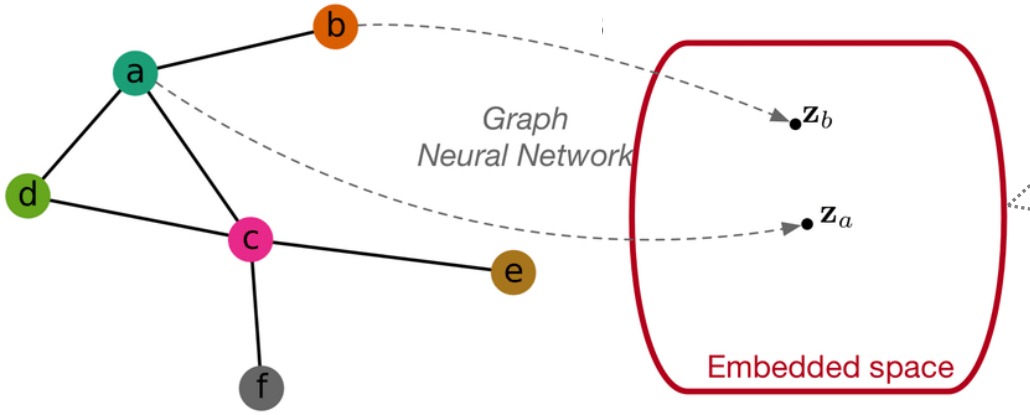
●: Blue Party
●: Green Party
●: ? Unknown

Prediction

●: Patient
●: Non-Patient
●: ? Unknown



Graph Neural Networks



Node classification:

$\mathrm{softmax}(\mathbf{z_n})$

e.g. Kipf & Welling (ICLR 2017)

Edge Prediction:

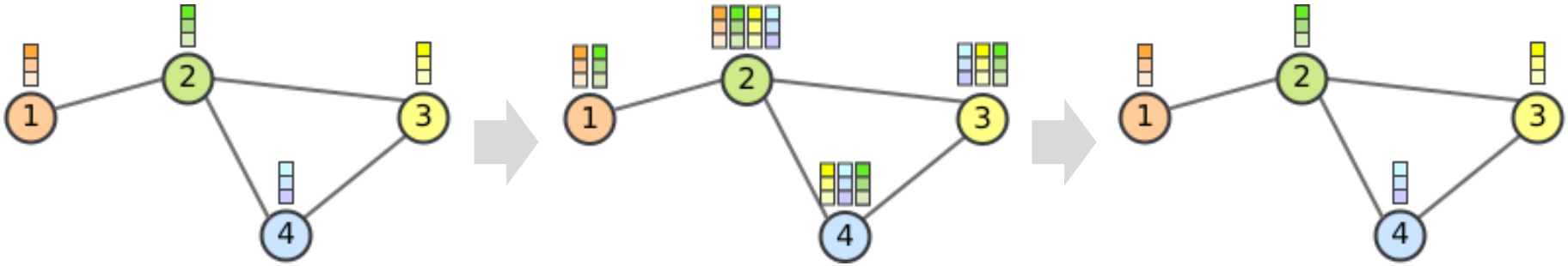
$p(A_{ij}) = \sigma(\mathbf{z_i^T z_j})$

Kipf & Welling (NIPS BDL 2016)

"Graph Auto-Encoders"

Message-Passing in GNNs

"A man is known by the company he keeps"



Neighborhood
Aggregation

Learnable
Combination

Combating Fakes & Frauds

	Detection	Robust Detection via GraphML
Fake News	GCAN [1] [ACL'20]	RERD [2] [TKDD'24]
Customs Frauds	DATE [3] [KDD'20]	GraphFC [4] [CIKM'23]
Food Frauds	Existing Supervised Methods	GraphCAR [5] [WWW'24]
Link Frauds		NetFense [6] [TKDE'23]

1. Lu and Li. "GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media." ACL 2020.
2. Zhuang et al. "Towards Robust Rumor Detection with Graph Contrastive and Curriculum Learning." ACM TKDD 2024.
3. Kim and Tsai et al. "DATE: Dual Attentive Tree-aware Embedding for Customs Frauds Detection." ACM KDD 2020.
4. Tsai et al. "GraphFC: Customs Fraud Detection with Label Scarcity." ACM CIKM 2023.
5. Yang et al. "Detecting Illicit Food Factories from Chemical Declaration Data via Graph-aware Self-supervised Contrastive Anomaly Ranking." ACM TheWebConf (WWW) 2024.
6. Hsieh and Li. "NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data." IEEE TKDE 2023.

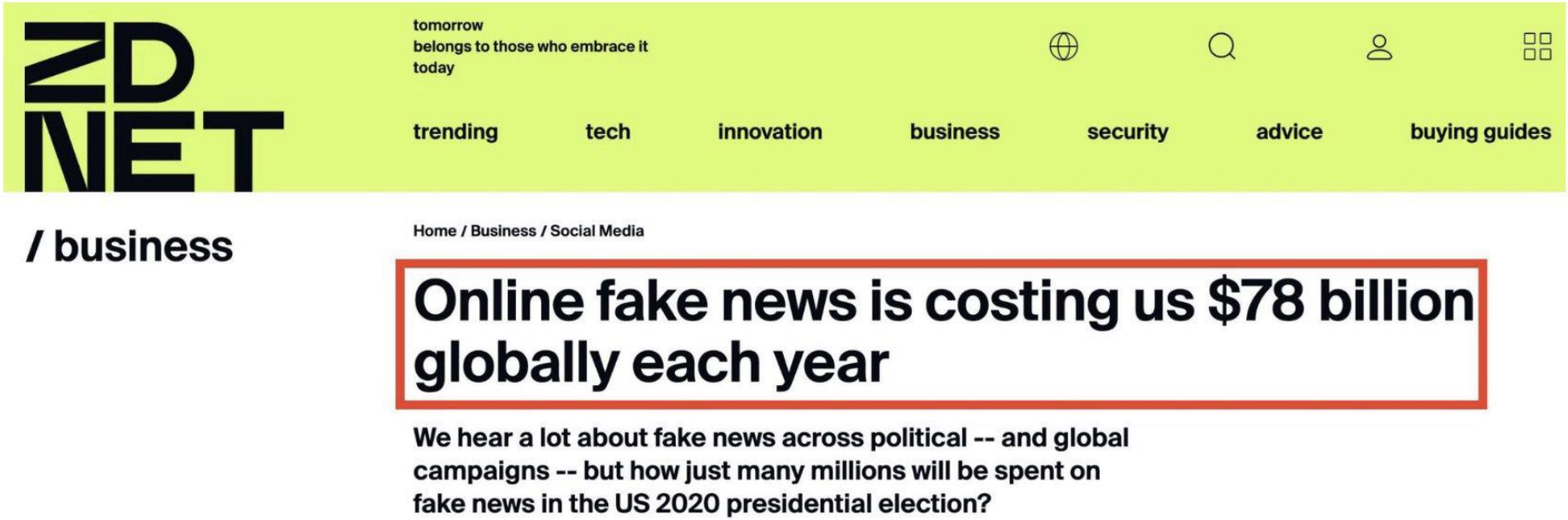
Combating Fakes & Frauds

	Detection	Robust Detection via GraphML
Fake News	GCAN [1] [ACL'20]	RERD [2] [TKDD'24]
Customs Frauds	DATE [3] [KDD'20]	GraphFC [4] [CIKM'23]
Food Frauds	Existing Supervised Methods	GraphCAR [5] [WWW'24]
Link Frauds		NetFense [6] [TKDE'23]

1. Lu and Li. "GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media." ACL 2020.
2. Zhuang et al. "Towards Robust Rumor Detection with Graph Contrastive and Curriculum Learning." ACM TKDD 2024.
3. Kim and Tsai et al. "DATE: Dual Attentive Tree-aware Embedding for Customs Frauds Detection." ACM KDD 2020.
4. Tsai et al. "GraphFC: Customs Fraud Detection with Label Scarcity." ACM CIKM 2023.
5. Yang et al. "Detecting Illicit Food Factories from Chemical Declaration Data via Graph-aware Self-supervised Contrastive Anomaly Ranking." ACM TheWebConf (WWW) 2024.
6. Hsieh and Li. "NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data." IEEE TKDE 2023.

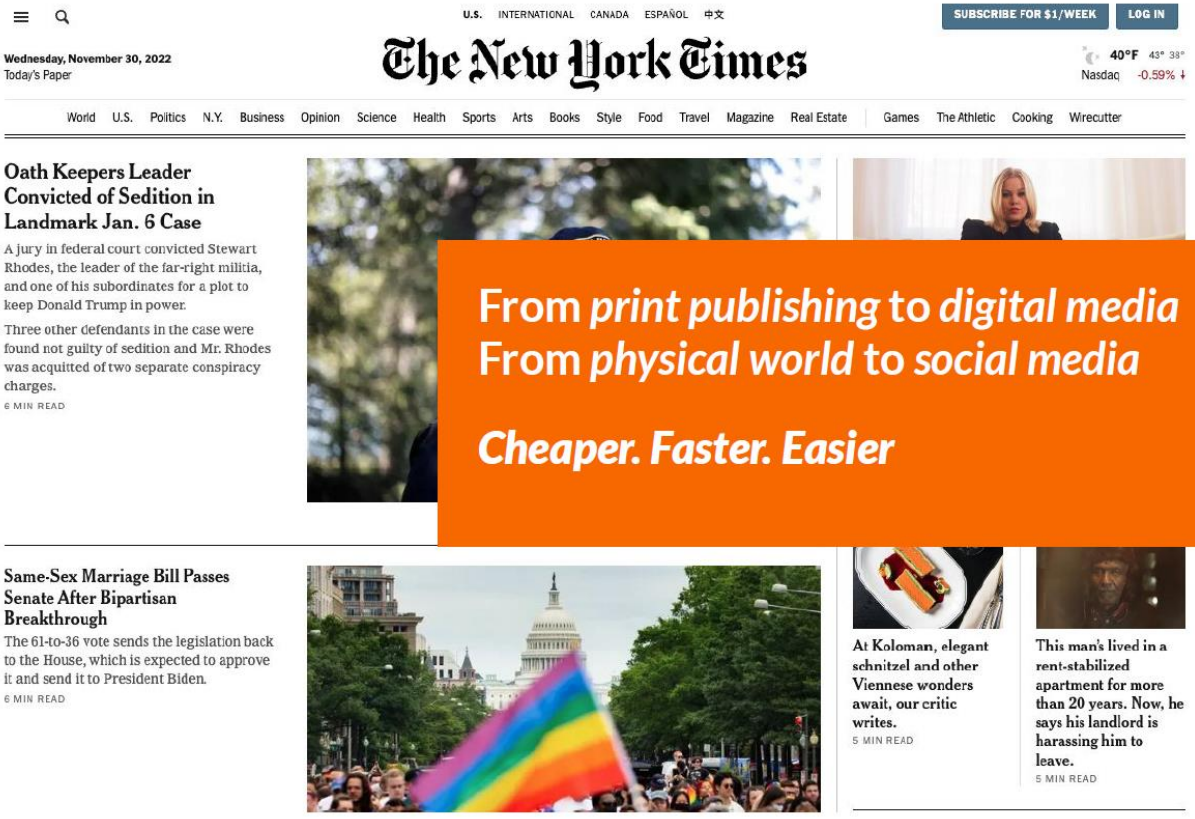
Why Fake News Research?

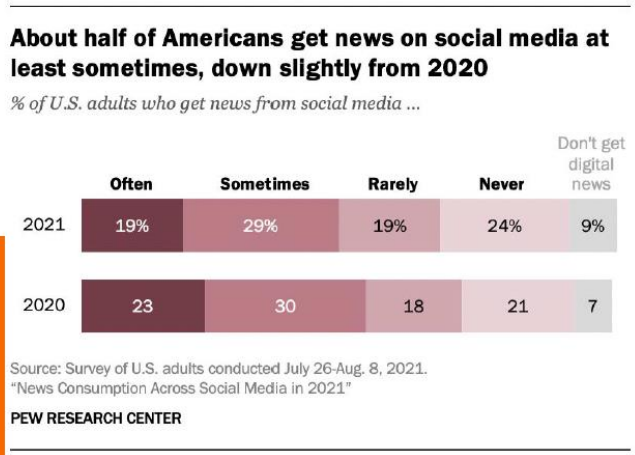
- Fake news is costing the global economy **$78 billion** each year [1]
- Social media is the **main source of news consumption** for younger generations [2]



[1] The Economic Cost of Bad Actors on the Internet https://s3.amazonaws.com/media.mediapost.com/uploads/EconomicCostOfFakeNews.pdf
[2] The news consumption habits of 16- to 40-year-olds https://www.americanpressinstitute.org/publications/reports/survey-research/the-news-consumption-habits-of-16-to-40-year-olds/

Fake News on Social Media



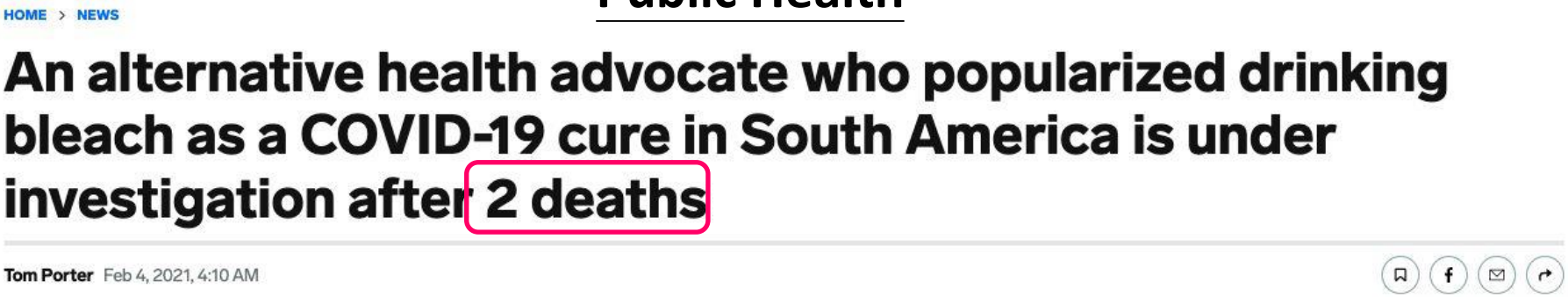


Why is combating fake news on social media essential?

Economics: wiped out $140 billion in stock value [1]

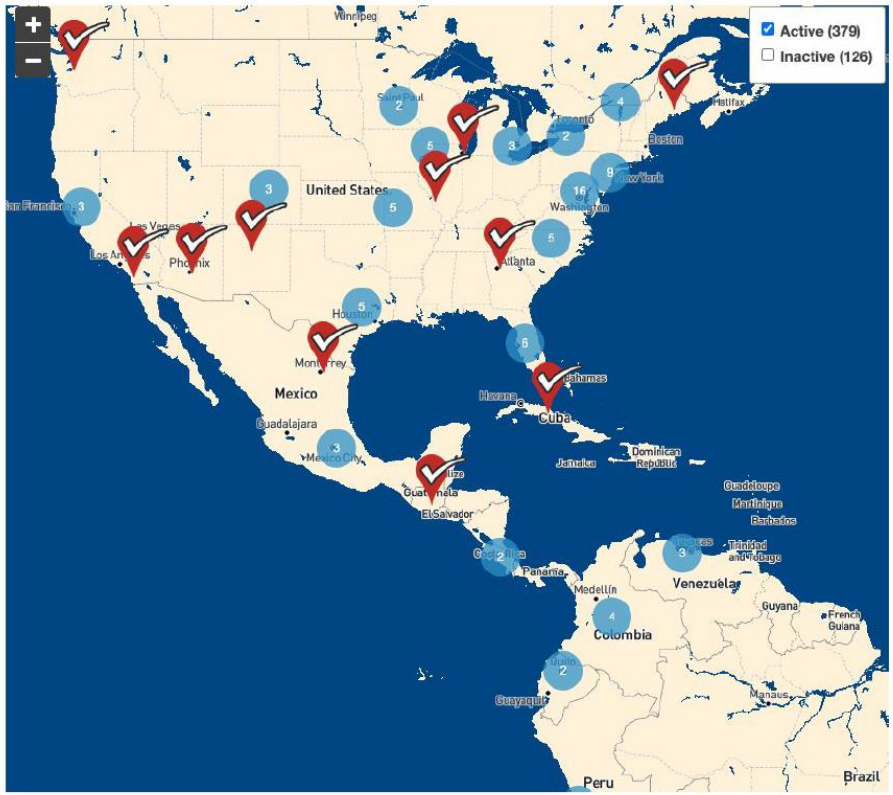
even though it was corrected in 7 mins

Public Health



[1] Johnson, S. C. (2013). Analysis: False white house tweet exposes instant trading dangers. Reuters.

Why is ML combating fake news on social media?



Duke Reporters' Lab

POLITIFACT
The Poynter Institute

LATEST PANTS ON FIRE! FACT-CHECKS

Instagram posts
stated on October 17, 2022 in an Instagram post:

Video shows Hillary Clinton talking about a glass dome over the Earth.

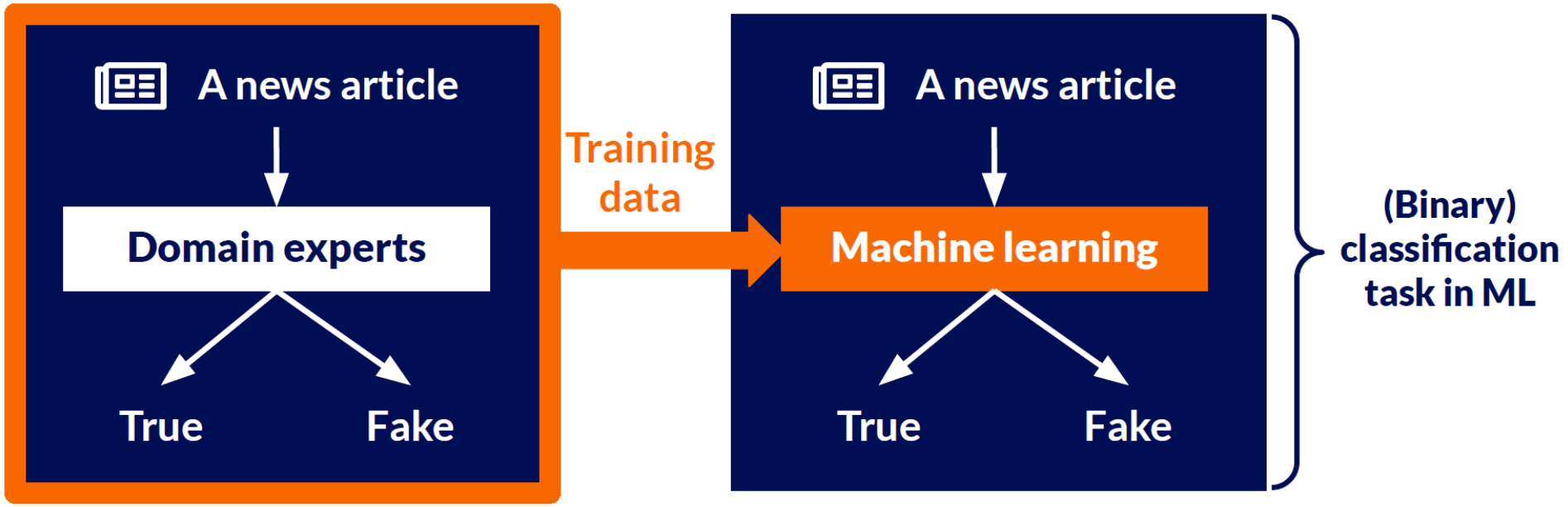
By Ciara O'Rourke • October 19, 2022

Viral image
stated on October 15, 2022 in an Instagram post:

Photo shows a taxi driver in New York City who "drives around taking people to kill."

By Ciara O'Rourke • October 19, 2022

Why is ML combating fake news on social media?

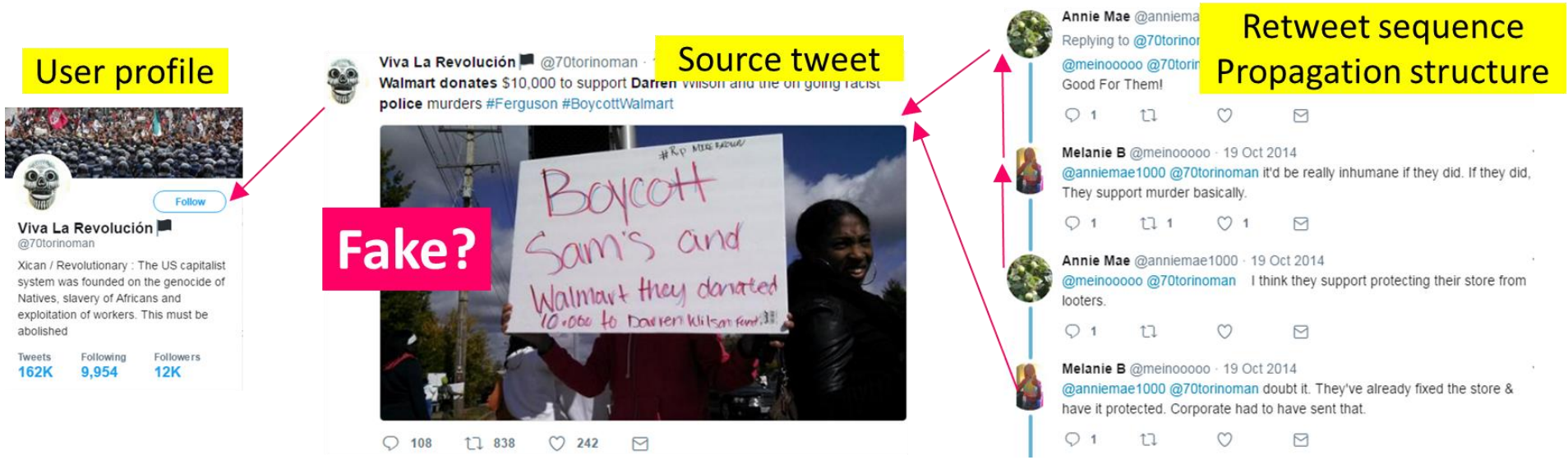


Manual fact-checking is **time-consuming** & **unscalable** though it often provides quality results

ML **automates** fact-checking, with manual labeling results as the ground-truth

Problem: Fake News Detection

- Given a source tweet (short **text**), along with
  - Retweet sequence or propagation structure (retweet, comment, like, etc.)
  - User profiles, and/or news media, URLs, image, context info (time, location, mention)
- Goals: **Predict whether the tweet is fake** **(binary classification)**



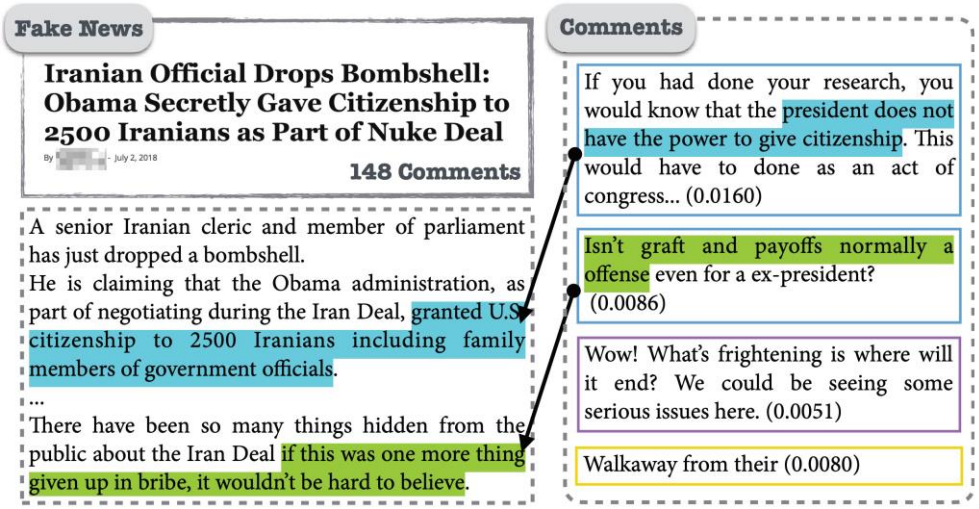
Detection in the **Wild**

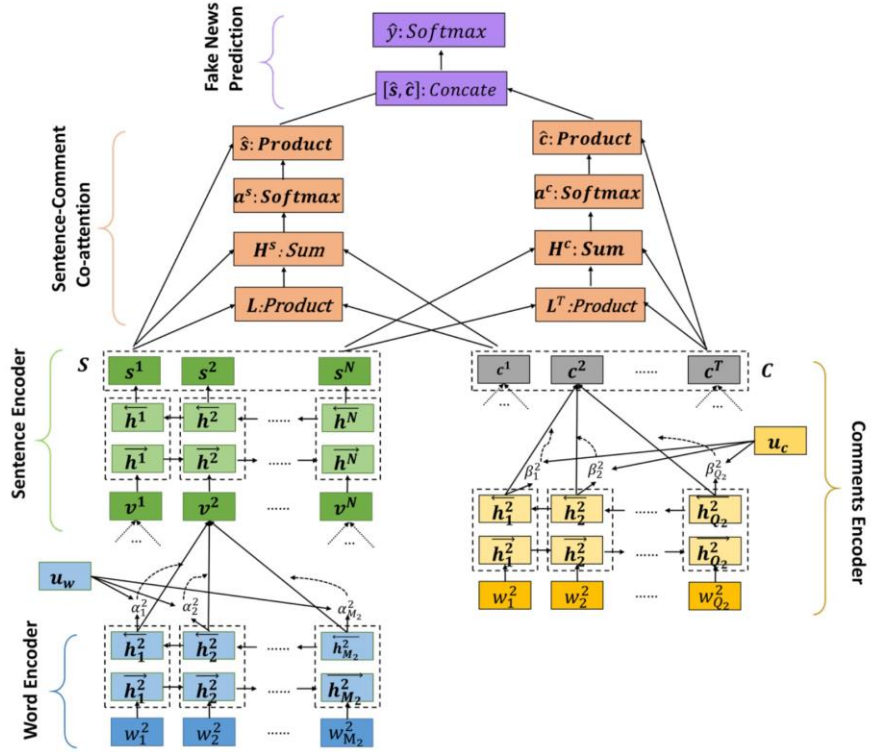
- Short texts with labels
- Leverage social contexts (e.g., propagation, user profiles)
- Accurate: high detection accuracy
- Early: early detection → detection before wide spread
- Limited labeled data: learning with label scarcity
- **Robust**
  - Against noisy data (text, propagation), and noisy labels
  - Against adversarial attack (on profile, text, propagation)
  - Against Large Language Models-generated texts

Detection in the **Wild**

Method	Venue	Text	Social Context	Early	Label Scarcity	Data Noise	Label Noise	Adver. Attacks
dEFNED	KDD'19	■		■				
Bi-GCN	AAAI'20	■	■					
GCAN	ACL'20	■	■	■				
FANG	CIKM'20	■	■		■			
DECOR	KDD'23	■		■	■	■		
GACL	WWW'22	■	■	■	■	■		
RDCL	CIKM'22	■	■		■	■		
MARL	WWW'23	■	■					■
UFNDA	Access'21	■		■	■ (no)			
RecMR	WWW'22	■	■	■	■ (no)			
REAL-FND	WWW'22	■	■		■ (cross)			
FactGen	AAAI'21	■			■			
InfoSurgeon	ACL'21	■			■			
RERD	TKDD'24	■	■	■	■	■	■	■

dEFEND: Explainable Fake News Detection

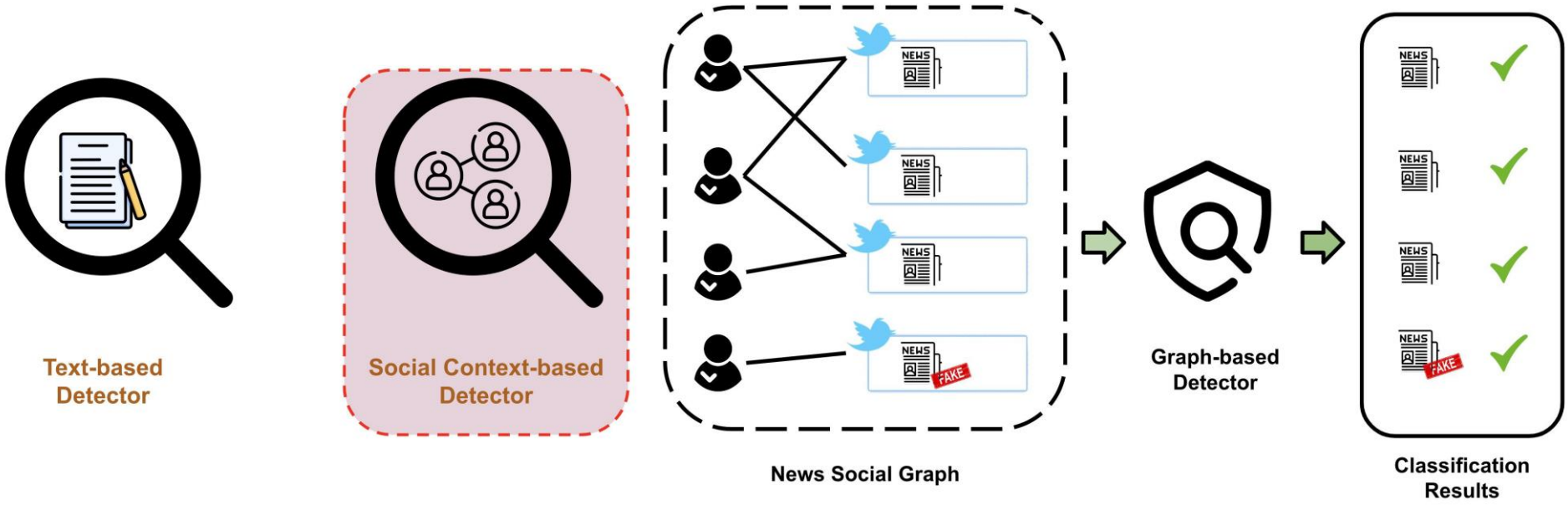




Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, Huan Liu. "dEFEND: Explainable Fake News Detection". KDD 2019

Are Existing Fake News Detectors **Robust**?

- Text-based detectors are **vulnerable** to adversarial attacks [1][2]
- What about **social context-based detectors**?

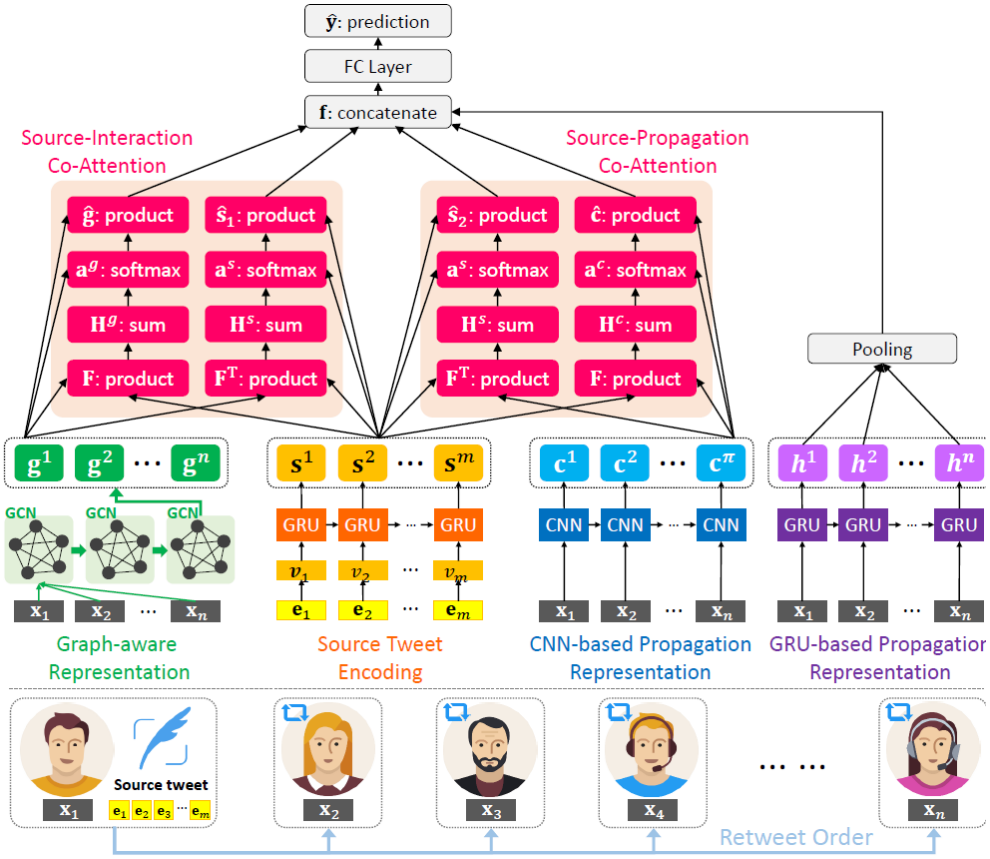


[1] He, Bing, Mustaque Ahamad, and Srijan Kumar. "Petgen: Personalized text generation attack on deep sequence embedding-based classification models." KDD 2021
[2] Le, Thai, Suhang Wang, and Dongwon Lee. "Malcom: Generating malicious comments to attack neural fake news detection models." IEEE ICDM 2020

Graph-aware Co-Attention Network (GCAN)

- No comments & propagation
- But aim to model user interactions
- Use GCN to learn user embeddings
- Co-attention between words & users

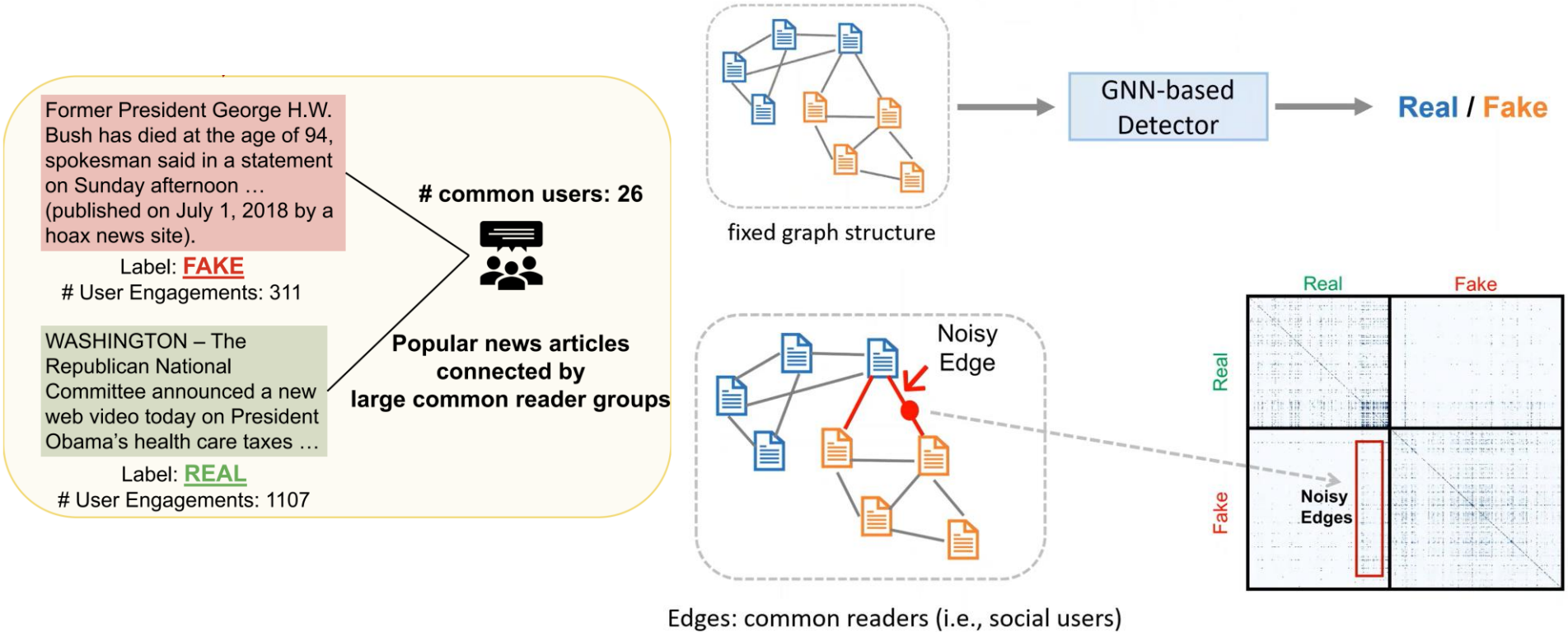
	Twitter15				Twitter16			
Method	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
DTC	0.4948	0.4806	0.4963	0.4949	0.5616	0.5369	0.5753	0.5612
SVM-TS	0.5190	0.5186	0.5195	0.5195	0.6915	0.6910	0.6928	0.6932
mGRU	0.5104	0.5148	0.5145	0.5547	0.5563	0.5618	0.5603	0.6612
RFC	0.4642	0.5302	0.5718	0.5385	0.6275	0.6587	0.7315	0.6620
tCNN	0.5140	0.5206	0.5199	0.5881	0.6200	0.6262	0.6248	0.7374
CRNN	0.5249	0.5305	0.5296	0.5919	0.6367	0.6433	0.6419	0.7576
CSI	0.7174	0.6867	0.6991	0.6987	0.6304	0.6309	0.6321	0.6612
dEFEND	0.6541	0.6611	0.6584	0.7383	0.6311	0.6384	0.6365	0.7016
GCAN-G	0.7938	0.7990	0.7959	0.8636	0.6754	0.6802	0.6785	0.7939
GCAN	**0.8250**	**0.8295**	**0.8257**	**0.8767**	**0.7593**	**0.7632**	**0.7594**	**0.9084**
Improvement	**15.0%**	**20.8%**	**18.1%**	**18.7%**	**19.3%**	**15.9%**	**3.8%**	**19.9%**



Y.-J. Lu, and C.-T. Li. "GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media." ACL 2020.

Shortcomings of Graph-based Approach

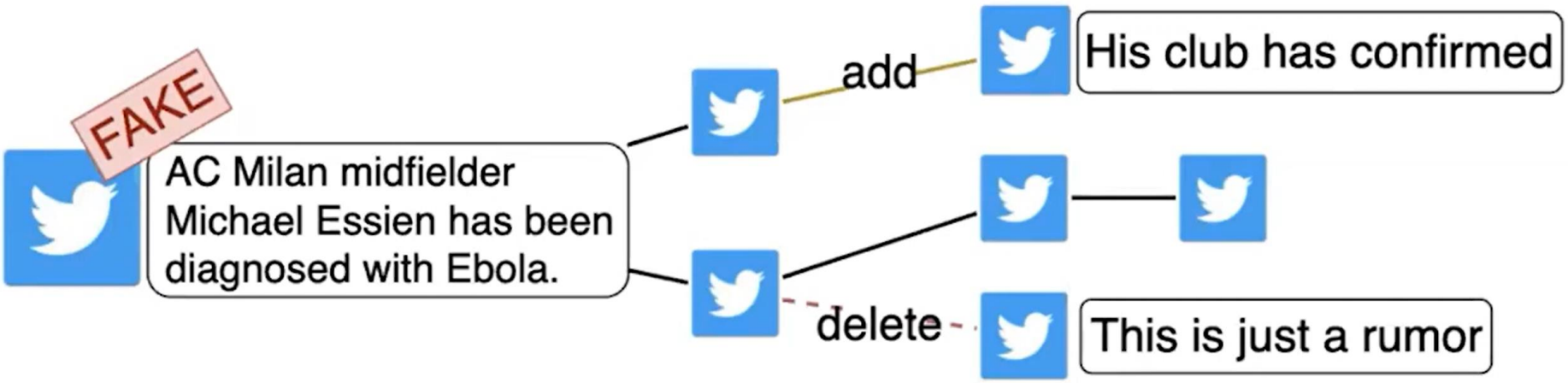
- Existing methods construct social graphs with varying features and capture structural patterns with GNNs
- However, noisy edges connecting influential real and fake news severely limit GNN effectiveness



"DECOR: Degree-Corrected Social Graph Refinement for Fake News Detection" KDD 2023

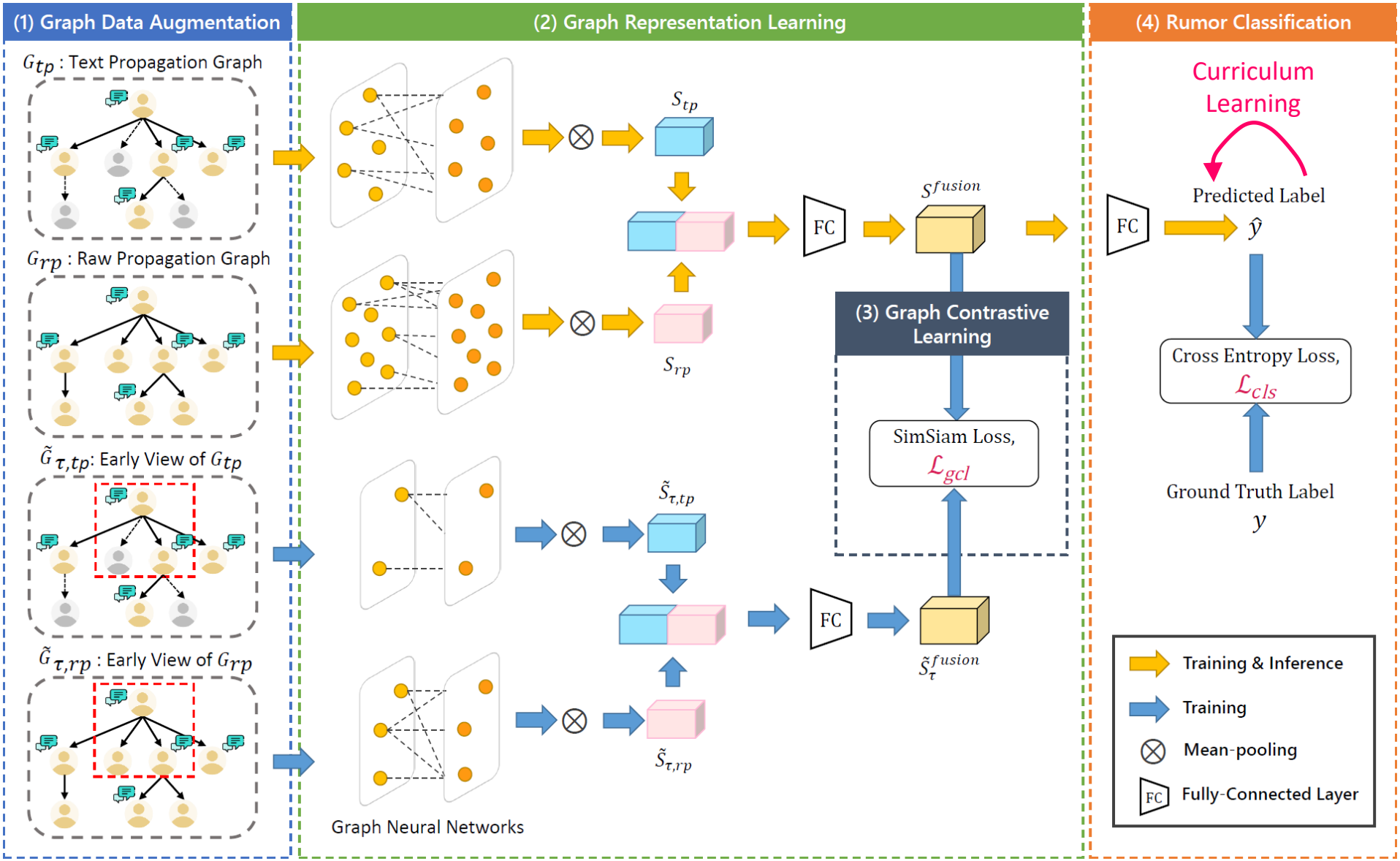
Adversarial Attacks on User Comments

- Lack robustness in face of noise and adversarial rumors, or even the conversational structures that is deliberately perturbed (e.g., adding or deleting some comments)



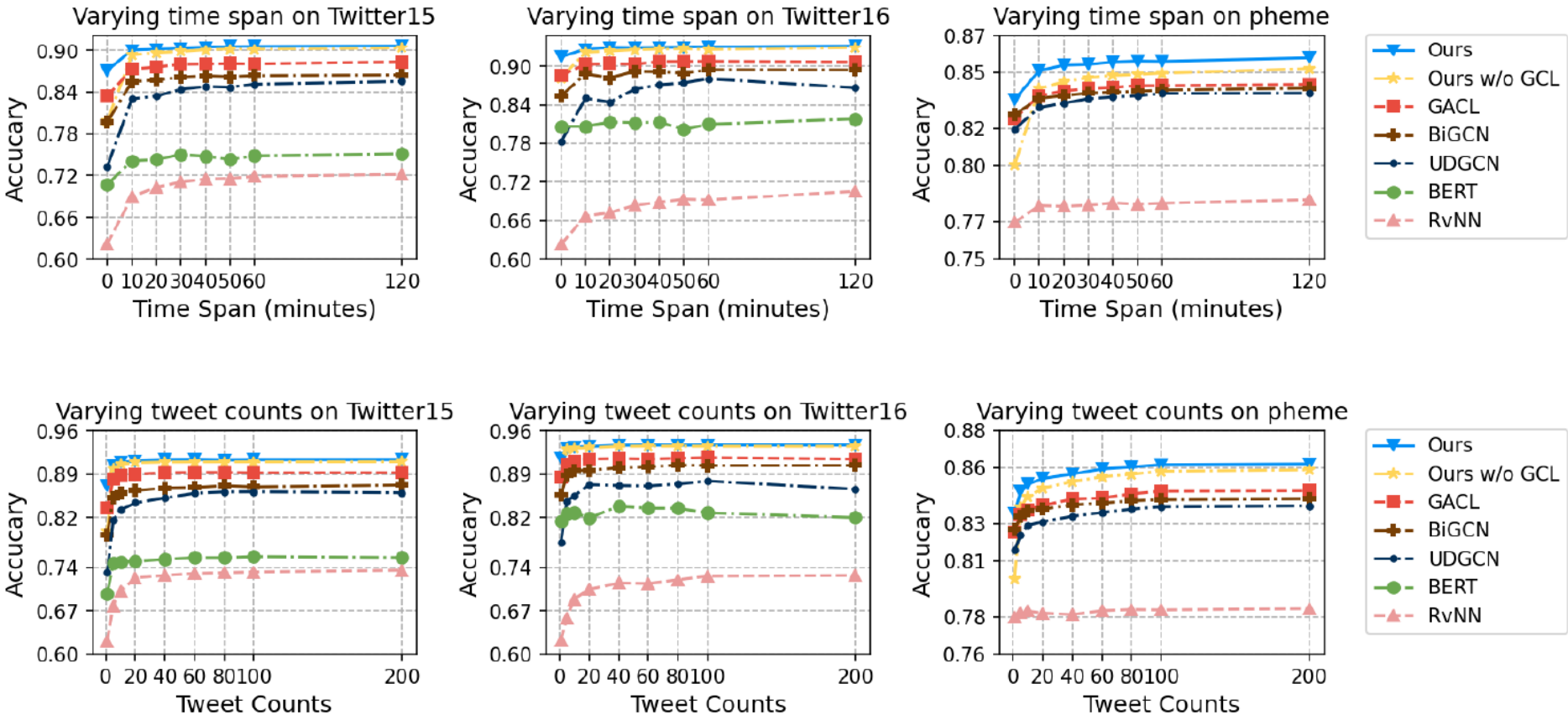
"Rumor Detection on Social Media with Graph Adversarial Contrastive Learning" TheWebConf (WWW) 2022

Robustness-Enhanced Rumor Detection (RERD)



Zhuang et al. "Towards Robust Rumor Detection with Graph Contrastive and Curriculum Learning." ACM TKDD 2024.

RERD on **Early** Detection



RERD on **Robust** Detection

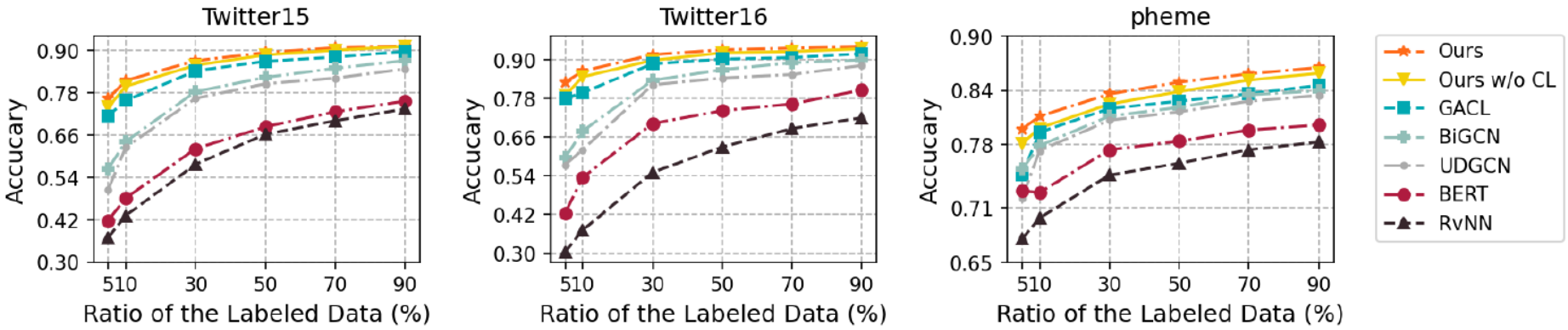


Fig. 5. Results for rumor detection under different label ratios.

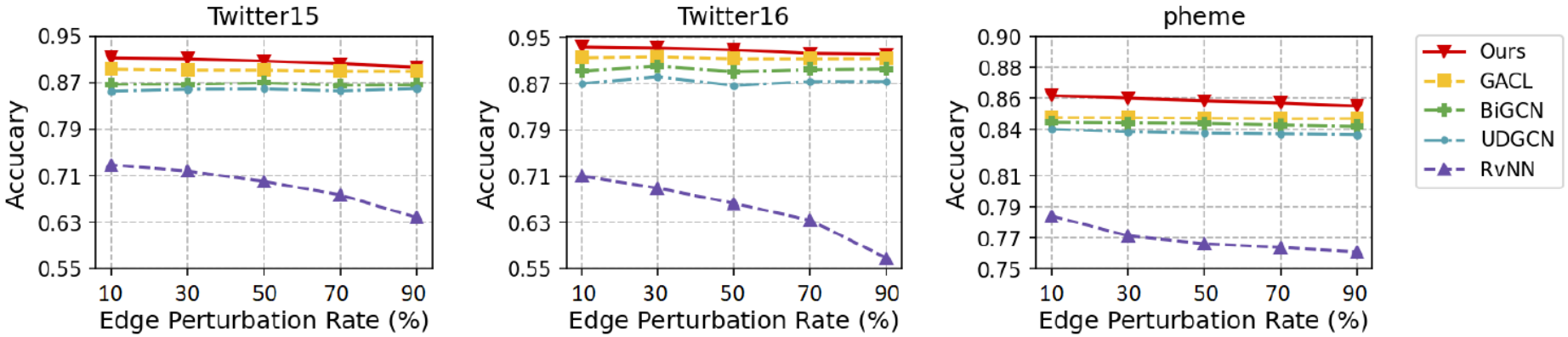


Fig. 6. Results for rumor detection under different edge perturbation ratios.

RERD on **Robust** Detection

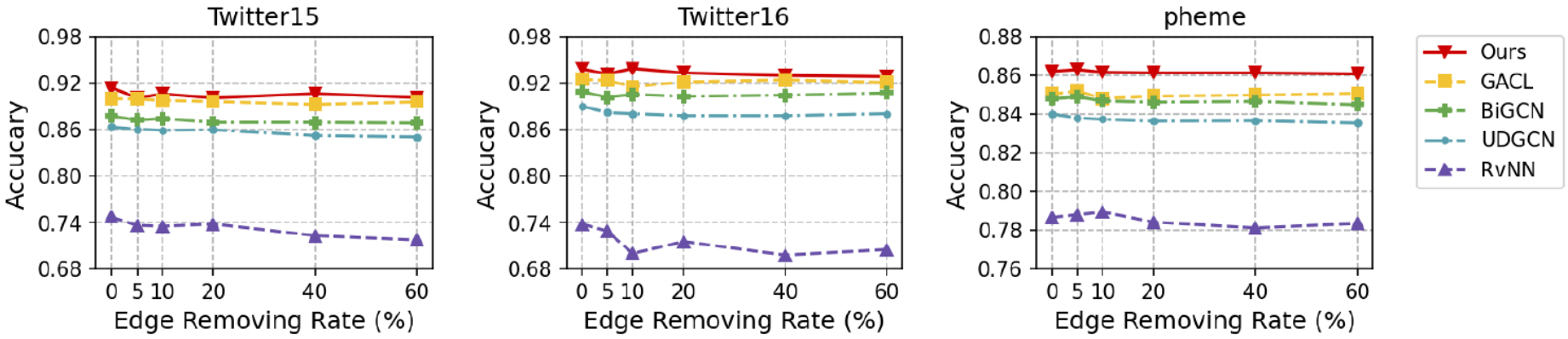


Fig. 7. Results for rumor detection under different edge removal ratios.

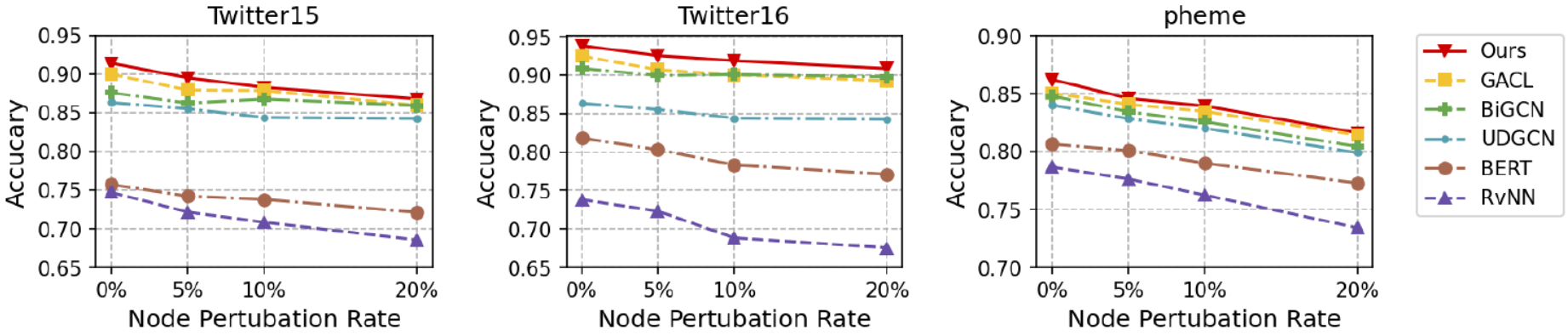


Fig. 8. Results for rumor detection under different label perturbation ratios.

Combating Fakes & Frauds

	Detection	Robust Detection via GraphML
Fake News	GCAN [1] [ACL'20]	RERD [2] [TKDD'24]
Customs Frauds	DATE [3] [KDD'20]	GraphFC [4] [CIKM'23]
Food Frauds	Existing Supervised Methods	GraphCAR [5] [WWW'24]
Link Frauds		NetFense [6] [TKDE'23]

1. Lu and Li. "GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media." ACL 2020.
2. Zhuang et al. "Towards Robust Rumor Detection with Graph Contrastive and Curriculum Learning." ACM TKDD 2024.
3. Kim and Tsai et al. "DATE: Dual Attentive Tree-aware Embedding for Customs Frauds Detection." ACM KDD 2020.
4. Tsai et al. "GraphFC: Customs Fraud Detection with Label Scarcity." ACM CIKM 2023.
5. Yang et al. "Detecting Illicit Food Factories from Chemical Declaration Data via Graph-aware Self-supervised Contrastive Anomaly Ranking." ACM TheWebConf (WWW) 2024.
6. Hsieh and Li. "NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data." IEEE TKDE 2023.

Common Illicit Items in Customs

Type of Frauds in Customs

Fraud Types	Illicit Motives	Our Scope
Undervaluation of trade goods	To avoid ad-valorem customs duty, or conceal illicit financial flows from exporters	Yes
Misclassification of HS code	To get a lower tariff rate applied or trade prohibited goods by avoiding restriction	
Manipulation of origin country	To get a preferential tariff rate under a free trade agreement	
Smuggling w/o declaration	To trade prohibited goods by avoiding restriction and customs duties	No
Overvaluation of trade goods	To disguise illicit financial flows as legitimate trade payment from importers	

TV (HS 852859, 8% duty)
PC Monitor (HS 852852, 0% duty)

Customs Fraud Detection

Import Declarations

ID	Item Category	Weight	Price	Importer	Origin Country
T1	Mushroom	35kg	250K	AAA	TH
T2	Wine	50kg	150K	BBB	UK
T3	Cigarette	90kg	200K	CCC	HK
T4	Face Mask	40kg	100K	DDD	JP
T5	Handbag	80kg	900K	EEE	FR
T6	Car Parts	90kg	120K	FFF	KR

Human Inspector

Illicit?
1
0
1
0
0
0

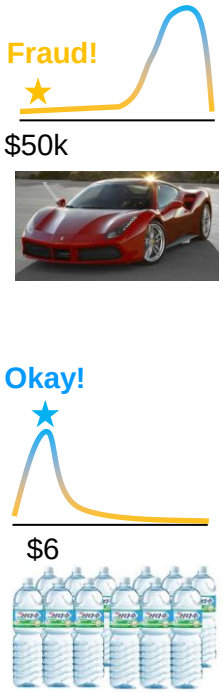
ID	Item Category	Weight	Price	Importer	Origin Country
T1	Mushroom	35kg	250K	AAA	TH
T2	Wine	50kg	150K	BBB	UK
T3	Cigarette	90kg	200K	CCC	HK
T4	Face Mask	40kg	100K	DDD	JP
T5	Handbag	80kg	900K	EEE	FR
T6	Car Parts	90kg	120K	FFF	KR

AI Model

Illicit?
0.75
0.60
0.90
0.05
0.30
0.01

Research Goals

- Given: import transaction t, importer u, HS code c
- Goals
  - Produce a ranking list of fraud transactions
  - Predict the raised revenue obtainable by inspecting a transaction



Given | **Prediction**

Item	Declaration	Importer	Predicted Illicit prob	Ranking (Priority)	Illicit (Truth)	Revenue (Truth)
Ferrari 488 Spider	$50,000	John	0.99	1	1	$90,000
Porsche 911	$30,000	Jane	0.98	2	1	$30,000
	...	...	...	...	...	...
Hyundai Elantra	$15,000	Kim	0.01	1000	0	$0
	...	...	...	...	...	...
Water 2Lx12	$6	Wang	0.00	10000	0	$0

Goal: Select 10% of items that requires inspection

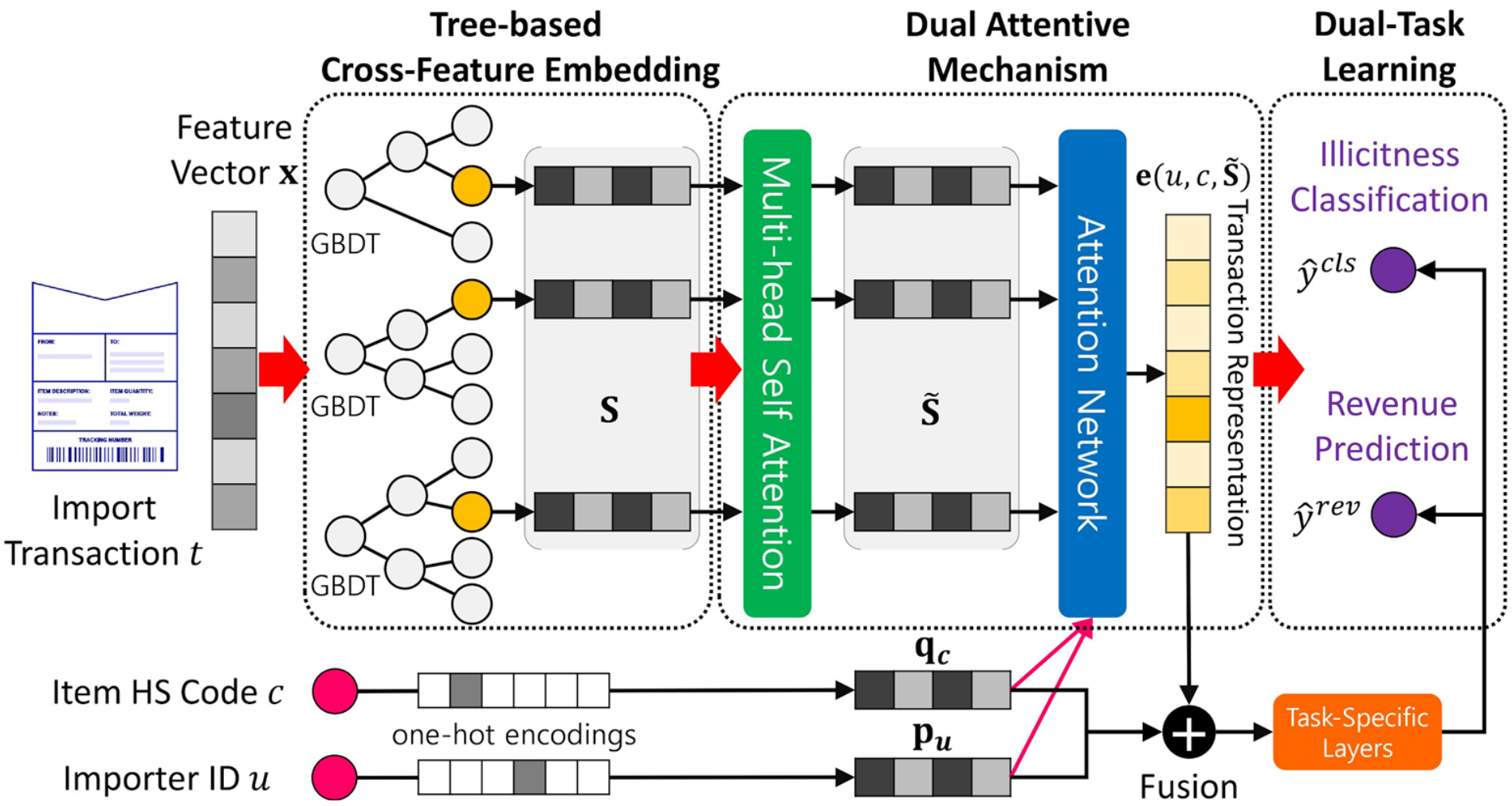
Customs Import Data

- 5-year import transactions from **Nigeria** Customs
- Since Nigeria had being doing 100% inspection on every import transaction, we are able to obtain the exact label for each transaction
  - Transactions with multiple items are removed
- Experimental settings:
  - Training: Y2013-2016 (1656K transactions) - 4.1% illicit rate
  - Testing: Y2017 (276K transactions) - 2.2% illicit rate

Detailed Data Fields

Type	Variable	Description	Example
Features	*sgd.id*	An individual numeric identifier for Single Goods Declaration (SGD).	SGD347276
	sgd.date	The year, month and day on which the transaction occurred.	13-11-28
	importer.id	An individual identifier by importer based on the tax identifier number (TIN) system.	IMP364856
	declarant.id	An individual identification number issued by Customs to brokers.	DEC795367
	country	Three-digit country ISO code corresponding to transaction.	USA
	office.id	The customs office where the transaction was processed.	OFFICE91
	tariff.code	A 10-digit code indicating the applicable tariff of the item based on the harmonised system (HS).	8703232926
	quantity	The specified number of items.	1
	gross.weight	The physical weight of the goods.	150kg
	fob.value	The value of the transaction excluding, insurance and freight costs.	$350
	cif.value	The value of the transaction including the insurance and freight costs.	$400
	total.taxes	Tariffs calculated by initial declaration.	$50
Prediction Target	***illicit***	Binary target variable that indicates whether the object has fraud.	1
	revenue	Amount of tariff raised after the inspection, only available on some illicit cases.	$20

DATE: Dual-Attentive Tree Embedding



S. Kim et al. "DATE: Dual Attentive Tree-aware Embedding for Customs Fraud Detection" ACM KDD 2020

Results – Comparison with Baselines

- Price/Importer: Select in the order of the declared price/fraud rate of importers
- IForest (Liu et al, 2008): Tree-based anomaly detection algorithm
- GBDT (Chen et al, 2016): XGBoost with cross features, trained on binary label y^{cls}
- GBDT+LR (He et al, 2014): Logistic regression based on cross features
- TEM (Wang et al, 2018): Tree-enhanced model with attentions
- **DATE_{CLS}, DATE_{REV}**: Custom selection with $\hat{y}^{cls}, \hat{y}^{rev}$ of DATE, respectively

	n = 1% (Selecting top 1%)			n = 2%			n = 5%			n = 10%			Overall	
Model	Pre.	Rec.	Rev.	Pre.	Rec.	Rev.	Pre.	Rec.	Rev.	Pre.	Rec.	Rev.	AUC	F1
Price	2.75%	1.23%	15.17%	2.23%	1.99%	20.64%	2.06%	4.60%	34.95%	2.30%	10.28%	50.98%	67.57%	7.81%
Importer	11.43%	5.10%	4.36%	9.41%	8.39%	7.56%	6.47%	14.43%	13.18%	5.22%	23.31%	30.31%	59.20%	9.10%
IForest	5.61%	2.50%	14.30%	6.19%	5.52%	23.14%	5.66%	12.62%	40.62%	5.12%	22.85%	54.14%	66.89%	5.28%
GBDT	90.01%	40.15%	24.59%	66.16%	59.04%	38.89%	32.19%	71.80%	57.20%	17.58%	78.42%	66.86%	93.38%	63.69%
GBDT+LR	90.95%	40.40%	27.18%	72.94%	65.09%	44.22%	35.02%	78.11%	63.77%	18.72%	83.54%	73.77%	94.82%	68.76%
TEM	88.72%	39.59%	39.48%	74.70%	66.43%	58.48%	37.39%	83.41%	78.58%	19.91%	88.54%	85.02%	96.52%	70.55%
DATE_{CLS}	**92.66%**	**41.33%**	44.97%	**80.79%**	**72.05%**	67.14%	**38.77%**	**86.49%**	84.35%	**20.24%**	**90.29%**	89.03%	**96.79%**	**75.32%**
DATE_{REV}	82.25%	36.63%	**49.29%**	79.93%	71.22%	**68.48%**	38.74%	86.41%	**84.57%**	20.11%	89.74%	**89.2%**	95.66%	75.23%

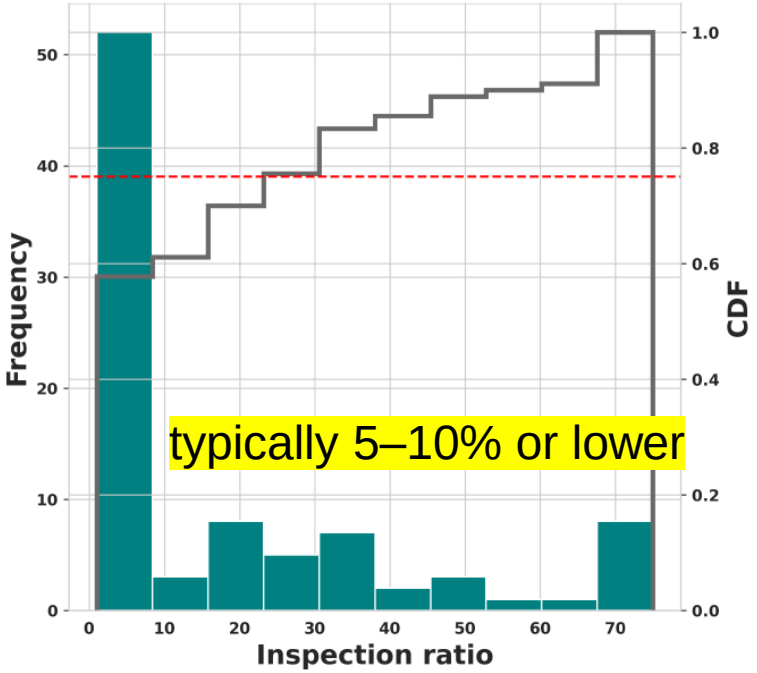
Results – Qualitative Studies

- To show that DATE has some potential to achieve **human-level interpretability**
- Select the top-2 significant cross features based on the highest attention scores

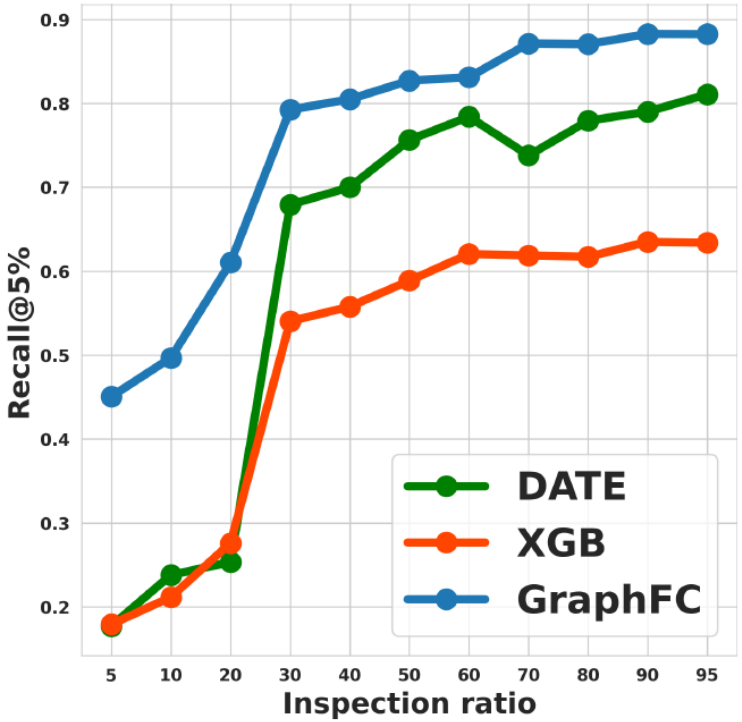
Comparison of illicit and licit transaction with respect to their corresponding cross features (CF) with highest attention score

	Illicit case	Licit case
Item	Used TOYOTA VENZA, $16,863	Used TOYOTA CAMRY, $4,673
CF 1	risk.importer=0 & tax.ratio<43.7% & gross.weight<3327.43 & fob.value>$1,366	12.2%<tax.ratio<16.8% & face.ratio>62.5%
CF 2	value/kg>$2 & cif.value>$1,912 & risk.(office,importer)=0 & tax.ratio <0.18%	risk.HS.origin=0 & value/kg<$2 & cif.value>$1,640 & risk.(office,importer)=0
$\hat{y}^{cls}$	0.9849	0.0001

Robustness Issue: Low Inspection Rate



(a) Distribution of physical inspection rate among 90 countries across the world.



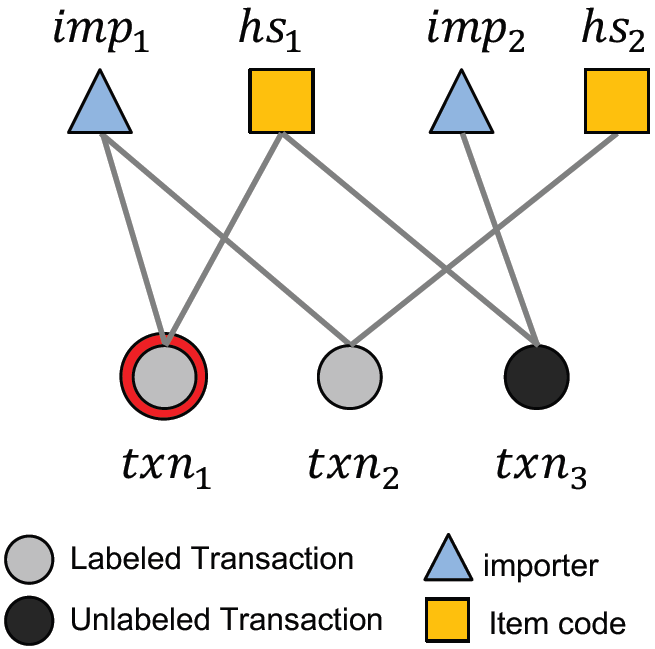
(b) Performance w.r.t. physical inspection ratio (%).

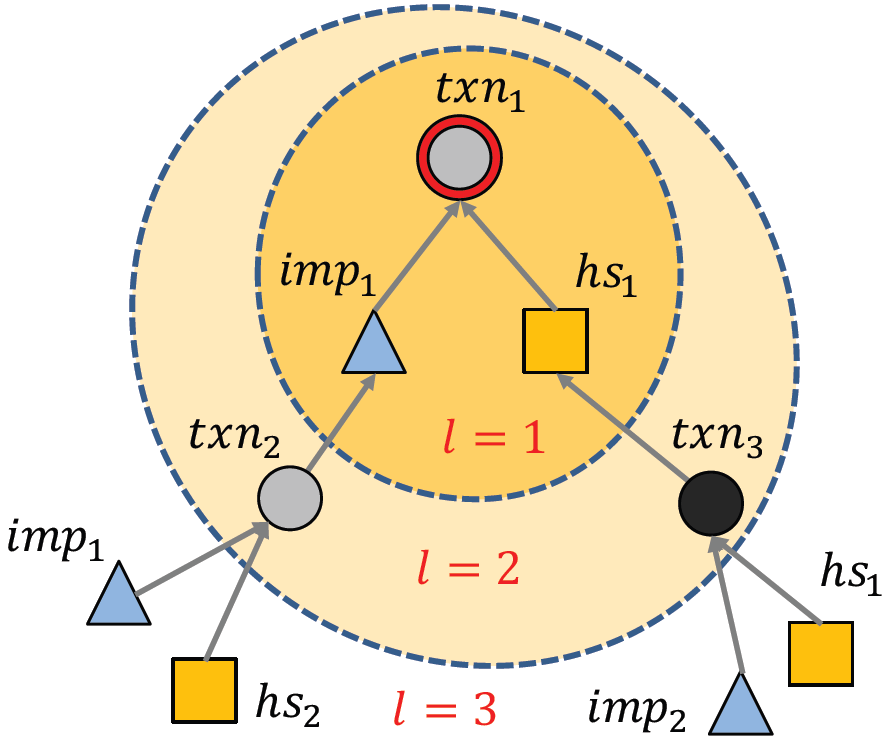
- Limited manpower and the colossal scale of trade volumes make manual inspection of all transactions practically impossible
- Low inspection rate → Lack of labeled data → Performance damage

Tackling the Challenges

- Issue 1: Limited labeled data
  - Make good use of rich unlabeled data
- Issue 2: Increasing unseen data
  - Need better generalization capability
- Our solution
  - Graph neural networks – semi-supervised learning
  - Propagate supervised signal
    from labeled to unlabeled data

Graph Formulation in Customs Declaration Data

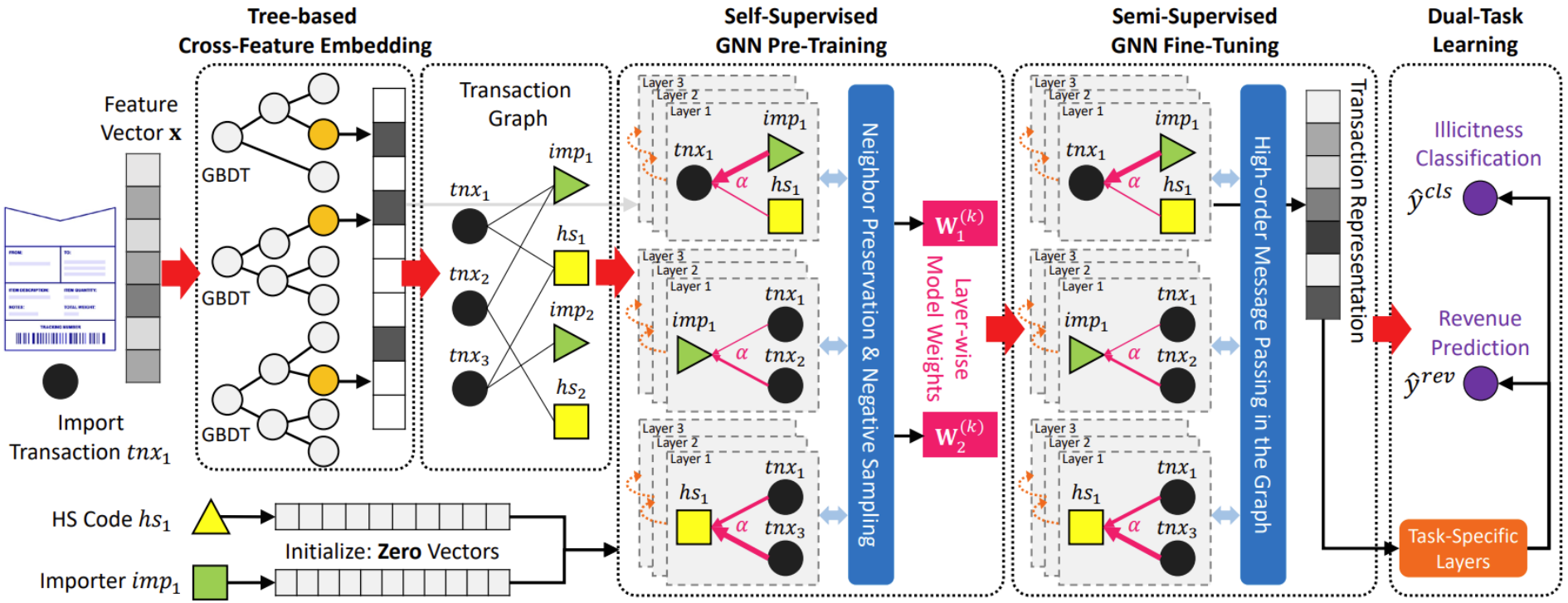




- Goal: classify the illicit labels of transaction nodes
  - Propagate unlabeled nodes' features to labeled nodes
  - Propagate labeled nodes' features to unlabeled nodes

GNN for Fraud Detection on Customs (GraphFC)

Step 1: Extract and utilize cross-features to initialize transaction node embeddings
Step 2: Self-supervised pre-training: make connected nodes' embeddings close
→ Align the distribution of unlabeled data closer to labeled data
→ Improve the generalization ability
Step 3 & 4: Semi-supervised fine-tuning with dual tasks



GraphFC Performance

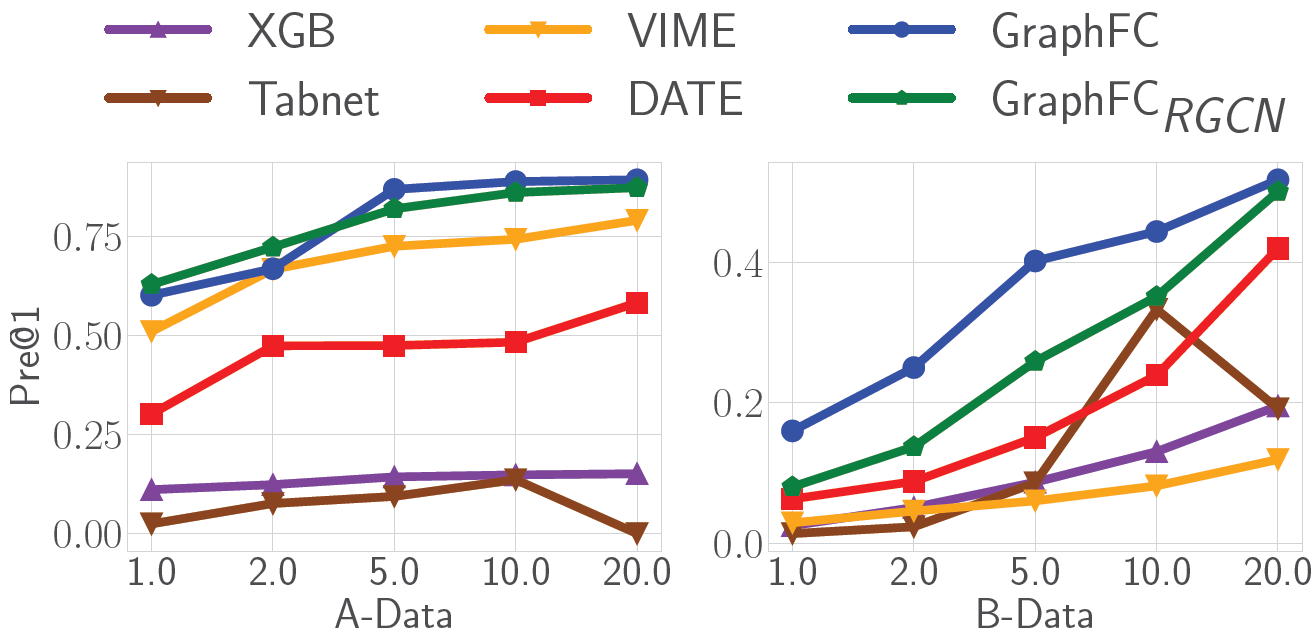
- **XGB**: XGBoost, the general winner for tabular data in Kaggle
- **DATE**: the SOTA method we just presented!
- **Tabnet**: the SOTA on tabular data classification with self-supervised learning
- **GraphFC$_{RGCN}$**: GNN with Relational GCN
- **GraphFC**: GNN with self-supervised pre-training and attention mechanism

Only 5% labeled data

A-Data						
Model	n=1%			n=5%		
	Pre.	Rec.	Rev.	Pre.	Rec.	Rev.
XGB	0.671	0.070	0.120	0.445	0.234	0.359
Tabnet	0.715	0.05	0.081	0.452	0.231	0.334
VIME	0.725	0.076	0.116	0.471	0.249	0.362
DATE	0.803	0.085	0.158	0.472	0.249	0.38
GraphFC$_{RGCN}$	0.819	0.086	0.146	0.525	0.277	0.424
GraphFC	**0.869**	**0.092**	**0.17**	**0.535**	**0.282**	**0.445**
B-Data						
XGB	0.151	0.061	0.109	0.045	0.092	0.184
Tabnet	0.149	0.06	0.089	0.046	0.085	0.171
VIME	0.064	0.024	0.052	0.043	0.087	0.208
DATE	0.152	0.061	0.105	0.057	0.115	0.210
GraphFC$_{RGCN}$	0.259	0.104	0.127	0.179	0.362	**0.334**
GraphFC	**0.402**	**0.162**	**0.172**	**0.200**	**0.406**	0.306

Performance by varying Inspection Rates

- Performance drops as the inspection rate decreases
- GraphFC still consistently outperform baselines
- GraphFC is robust against different degrees of label scarcity, and generalize well in different countries



Combating Fakes & Frauds

	Detection	Robust Detection via GraphML
Fake News	GCAN [1] [ACL'20]	RERD [2] [TKDD'24]
Customs Frauds	DATE [3] [KDD'20]	GraphFC [4] [CIKM'23]
Food Frauds	Existing Supervised Methods	GraphCAR [5] [WWW'24]
Link Frauds		NetFense [6] [TKDE'23]

1. Lu and Li. "GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media." ACL 2020.
2. Zhuang et al. "Towards Robust Rumor Detection with Graph Contrastive and Curriculum Learning." ACM TKDD 2024.
3. Kim and Tsai et al. "DATE: Dual Attentive Tree-aware Embedding for Customs Frauds Detection." ACM KDD 2020.
4. Tsai et al. "GraphFC: Customs Fraud Detection with Label Scarcity." ACM CIKM 2023.
5. Yang et al. "Detecting Illicit Food Factories from Chemical Declaration Data via Graph-aware Self-supervised Contrastive Anomaly Ranking." ACM TheWebConf (WWW) 2024.
6. Hsieh and Li. "NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data." IEEE TKDE 2023.

Food Safety Issues

Toxic PFAS chemicals used in packaging can end up in food, study finds

Compostable packaging is popular for environmental reasons, but it can be treated with 'forever chemicals' linked to health problems

Paper and cardboard food packaging treated with PFAS chemicals to make it grease-proof has concerned experts. Photograph: AAP

https://www.theguardian.com/environment/2023/apr/17/pfas-forever-chemicals-food-containers-study

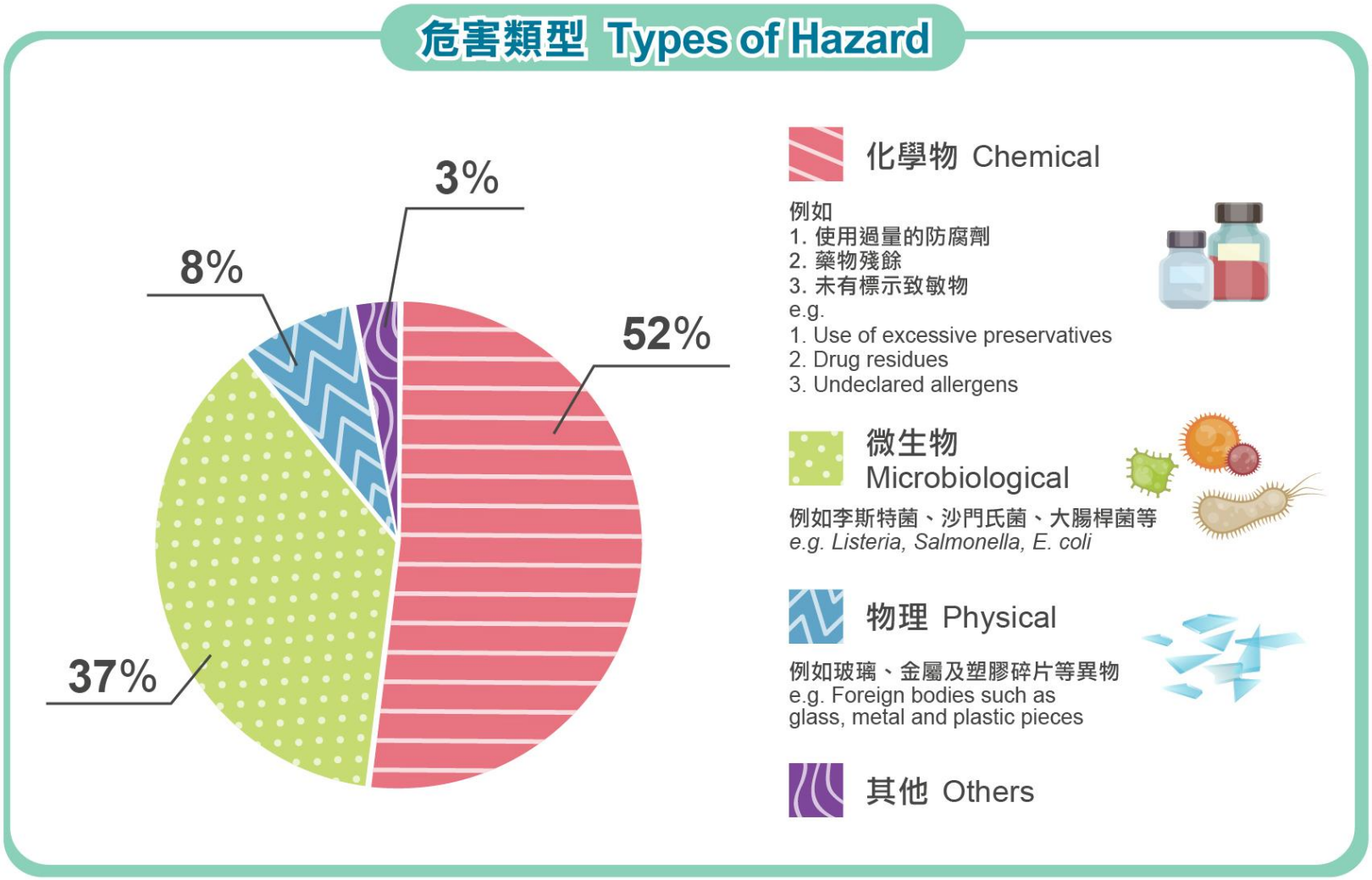
euronews.health HEALTH HEALTHCARE NUTRITION WELLBEING SERIES

Health > Health news

Nitrosamines: What to know about the cancer-causing chemicals being found in our food and medicines

https://www.euronews.com/health/2023/06/02/nitrosamines-what-to-know-about-the-cancer-causing-chemicals-being-found-in-our-food-and-m

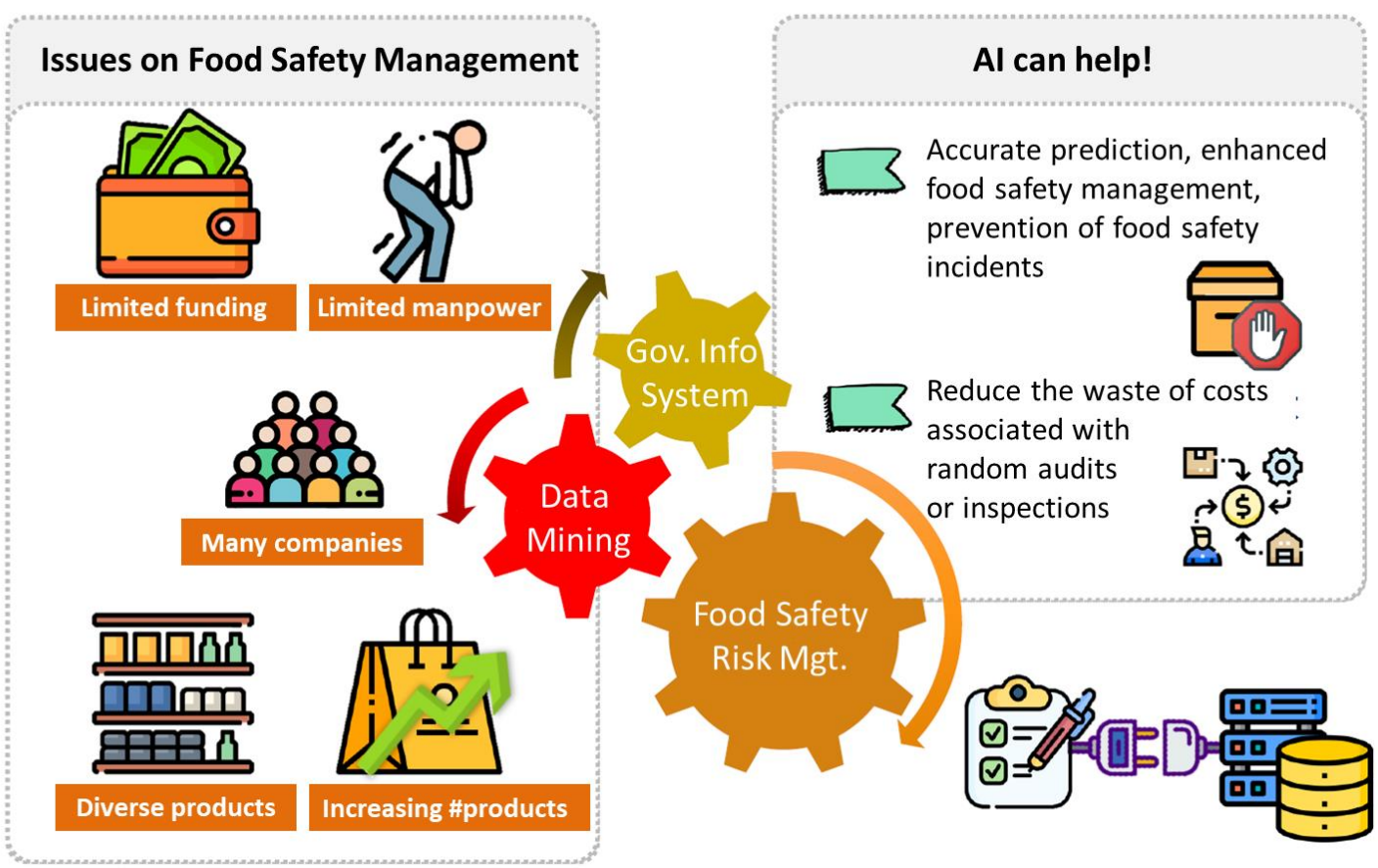
Chemical Hazards in Food Safety



https://www.cfs.gov.hk/english/multimedia/multimedia_pub/multimedia_pub_fsf_178_01.html

Challenges – AI can help!

Human Inspection



AI's Role: Enhancing Human Inspection!

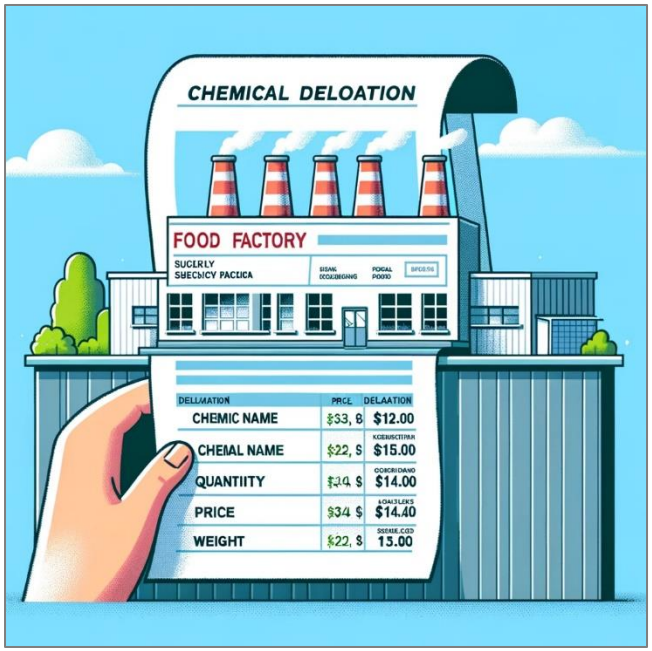
Utilize AI to identify potential illicit food factories.

Human experts conduct physical inspections at the AI-identified factories, where harmful chemicals could be added during food production.

Images are generated by DALL.E

Chemical Declaration Data

- In collaboration with local authorities
  - Digital declaration
  - On-site inspection
- 149 inspected food factories
  - 36 are identified as illicit (24%)
- 5,081 unique chemicals
- 58,667 chemical declaration records
  - Chemical name
  - Price
  - Weight
  - Quantity



Unsatisfying Supervised Detection

- Given food factories, each is associated with a set of declared chemicals and a binary label → Learn a binary classifier

	Train/Test = 80/20			Train/Test = 20/80		
	AUC	F1	AP	AUC	F1	AP
RF	59.66	33.76	34.50	52.50	12.43	27.21
LR	62.87	40.59	38.45	52.82	14.21	27.18
MLP	65.36	45.09	40.33	53.09	16.98	27.03
XGBoost	54.54	22.23	30.28	50.44	6.21	25.11
LightGBM	58.81	31.26	33.37	51.30	7.79	26.10
CatBoost	58.18	29.31	35.03	52.01	9.11	26.45

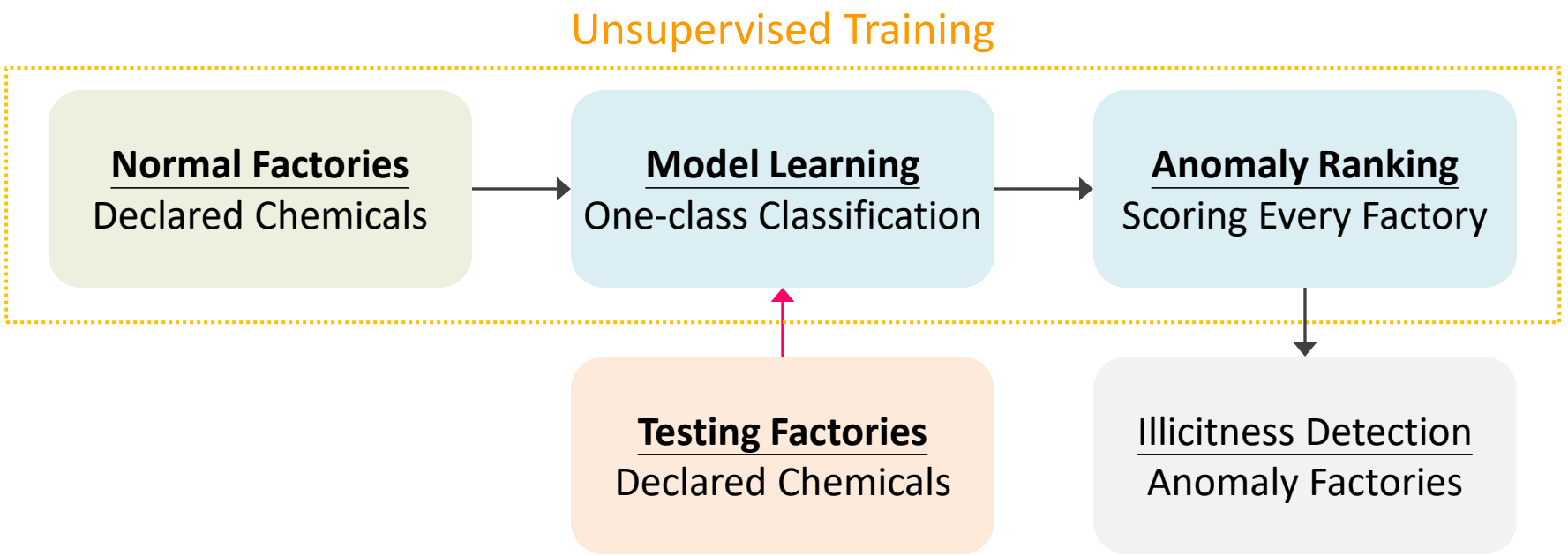
- Why unsatisfying? → limited #factory, class imbalance

Challenges: The Robustness Issues

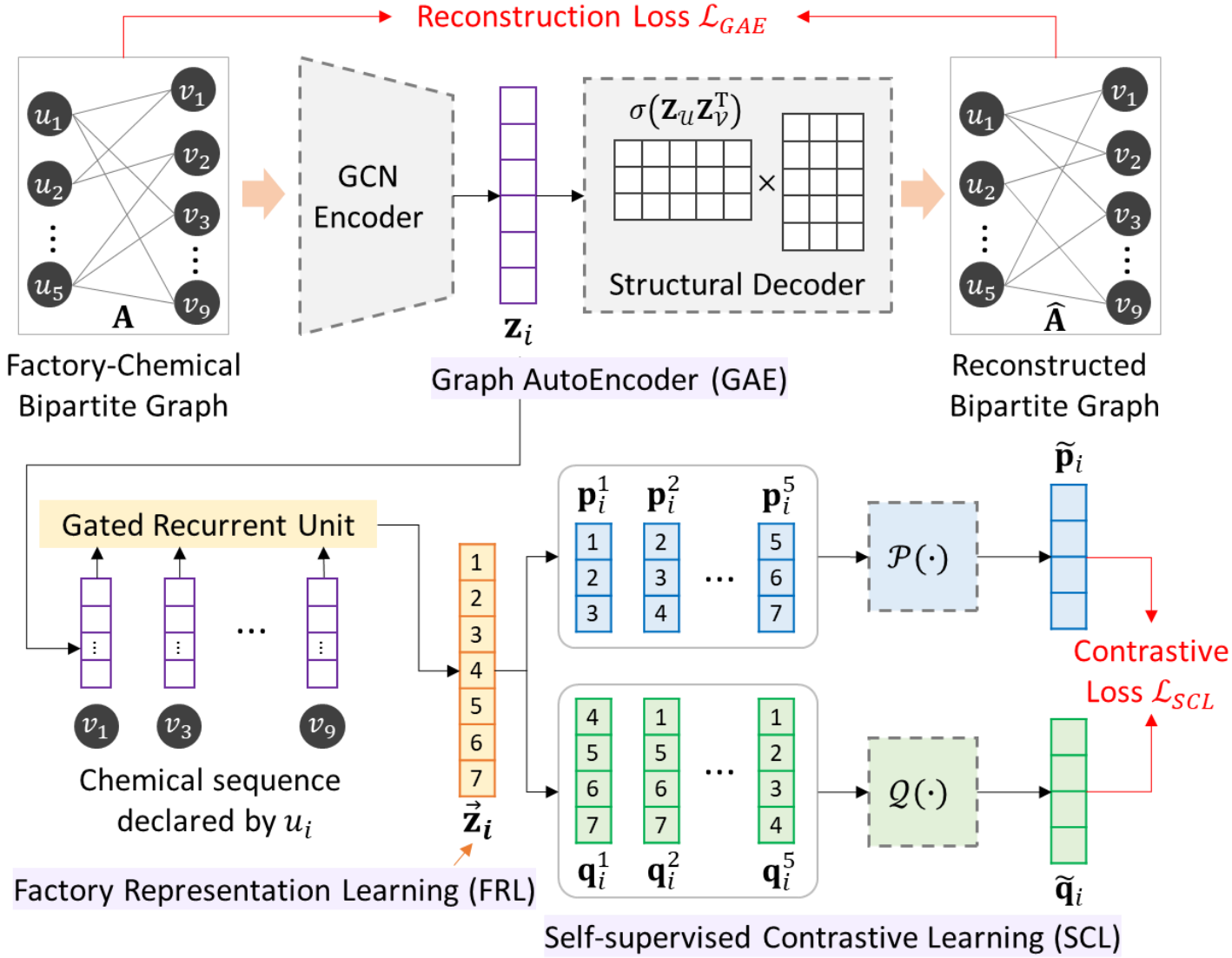
1) **Lack of Labeled Data**
   - Only a small % of factories are identified as illicit
2) **Vast Range of Declared Chemicals**
   - A sheer number of chemicals can be declared
3) **Noisy and Complicated Chemical Data**
   - Misspellings, naming variations, synonyms
4) **Subtle Illicit Evidences**
   - Illicit can stem from combinations of licit ones

Goal: Unsupervised Learning for Robust Detection

- Given food factories, each is associated with a set of declared chemicals, and labels are absent
- Goal: learning to score factories for anomaly ranking



Graph-aware self-supervised **C**ontrastive **A**nomaly **R**anking (**GraphCAR**)



Promising Outcome

- GraphCAR consistently outperforms all SOTA unsupervised anomaly detection models across metrics and training settings!

		AUC					F1					AP				
	Training %	90%	70%	50%	30%	10%	90%	70%	50%	30%	10%	90%	70%	50%	30%	10%
UOD	kNN [1]	51.93	51.47	50.89	49.30	43.83	47.80	47.71	47.71	47.71	48.98	31.41	31.33	31.33	31.33	33.33
	LOF [4]	51.18	50.73	50.92	49.43	43.93	47.71	47.71	47.71	47.71	47.71	31.33	31.33	31.33	31.33	31.33
	iForest [34]	41.81	44.78	44.08	36.16	33.33	47.71	47.71	47.71	47.71	47.71	31.33	31.33	31.33	31.33	31.33
	SO-GAAL [36]	51.34	53.09	49.20	46.24	59.04	48.21	49.59	48.51	**50.32**	61.18	31.98	35.46	32.02	33.62	44.07
	COPOD [32]	36.94	38.15	41.00	40.25	43.25	47.89	48.57	49.52	48.75	53.93	31.53	32.31	32.91	32.23	38.10
	ECOD [33]	36.99	38.15	41.19	40.27	42.98	47.89	48.57	49.52	48.75	53.76	31.53	32.31	32.91	32.23	37.31
	LUNAR [17]	51.24	50.49	49.21	50.74	46.96	47.96	48.28	48.15	48.51	51.61	32.34	31.82	31.71	32.74	35.82
	DIF [62]	53.98	49.09	47.97	46.47	57.36	47.71	48.14	48.96	49.06	53.76	31.33	32.11	32.49	32.50	37.31
OCC	OCSVM [52]	51.69	51.48	50.51	48.64	42.68	47.80	47.90	47.88	47.85	48.98	31.41	31.49	31.48	31.45	33.33
	VAE [28]	49.40	48.83	50.69	43.96	39.20	47.80	47.77	48.39	47.92	52.00	31.41	31.38	32.08	31.92	35.14
	β-VAE [20]	50.06	36.91	52.70	42.03	46.83	49.62	49.05	49.71	50.16	51.61	33.14	32.55	33.16	33.48	35.82
	OC4Seq [61]	60.76	62.97	62.96	54.35	47.30	48.70	49.42	49.88	48.25	48.15	33.13	33.67	36.24	32.07	31.71
	GraphCAR	**74.74**	**76.18**	**74.19**	**74.74**	**73.28**	**54.70**	**57.14**	**53.85**	50.00	**65.39**	**56.30**	**54.31**	**51.12**	**48.86**	**50.83**

Combating Fakes & Frauds

	Detection	Robust Detection via GraphML
Fake News	GCAN [1] [ACL'20]	RERD [2] [TKDD'24]
Customs Frauds	DATE [3] [KDD'20]	GraphFC [4] [CIKM'23]
Food Frauds	Existing Supervised Methods	GraphCAR [5] [WWW'24]
Link Frauds		NetFense [6] [TKDE'23]

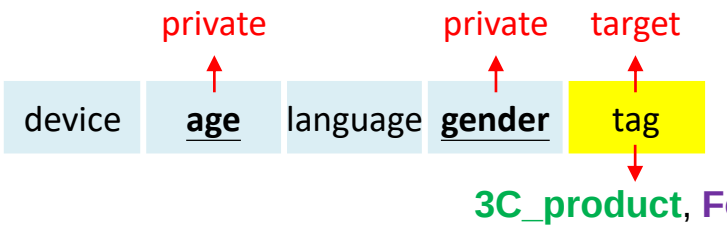
1. Lu and Li. "GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media." ACL 2020.
2. Zhuang et al. "Towards Robust Rumor Detection with Graph Contrastive and Curriculum Learning." ACM TKDD 2024.
3. Kim and Tsai et al. "DATE: Dual Attentive Tree-aware Embedding for Customs Frauds Detection." ACM KDD 2020.
4. Tsai et al. "GraphFC: Customs Fraud Detection with Label Scarcity." ACM CIKM 2023.
5. Yang et al. "Detecting Illicit Food Factories from Chemical Declaration Data via Graph-aware Self-supervised Contrastive Anomaly Ranking." ACM TheWebConf (WWW) 2024.
6. Hsieh and Li. "NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data." IEEE TKDE 2023.

AI + Social Media = Privacy Leaking

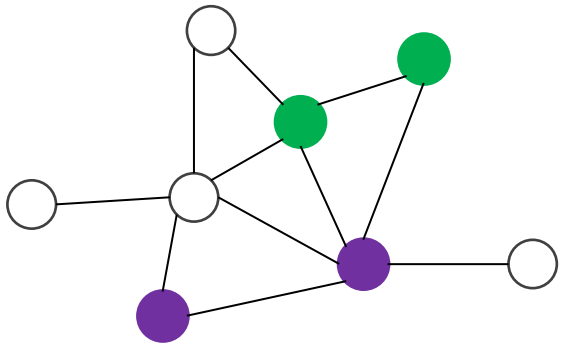
Node classification can be used to predict user private labels (e.g., gender, age)

https://www.protocol.com/bulletins/instagram-chronological-feed

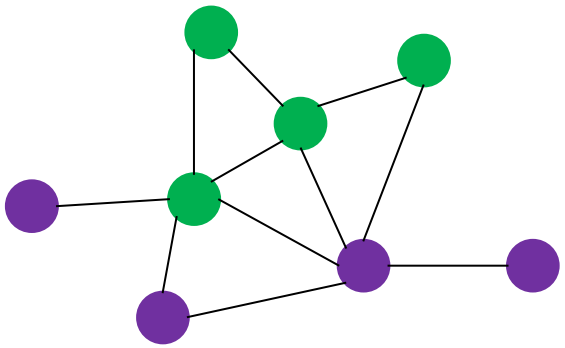
Link Frauds via Adversarial Attack on Graph Data



Target Label Classification (TLC)



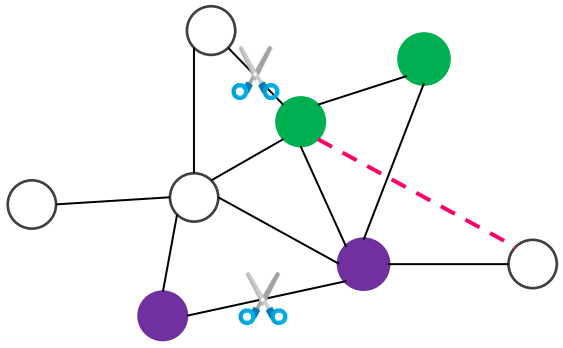
GNN



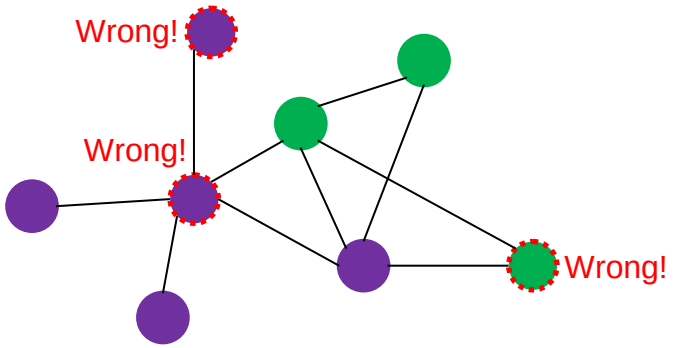
Adversarial Attacks

Link Frauds

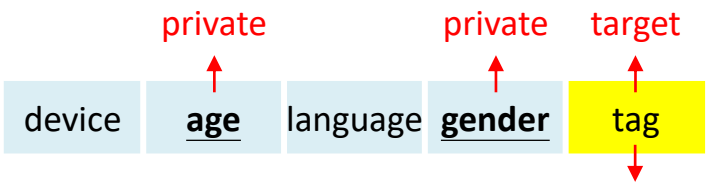
Adversarial Attacks on TLC



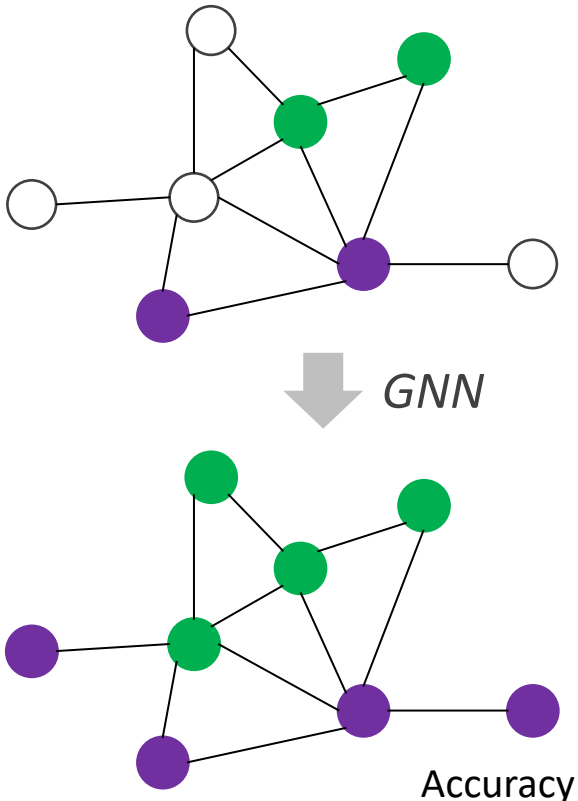
GNN

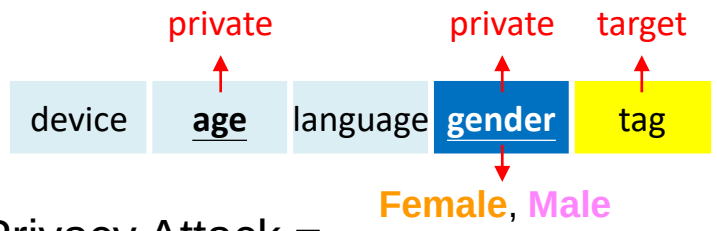


Injecting Link Frauds for Privacy Attack



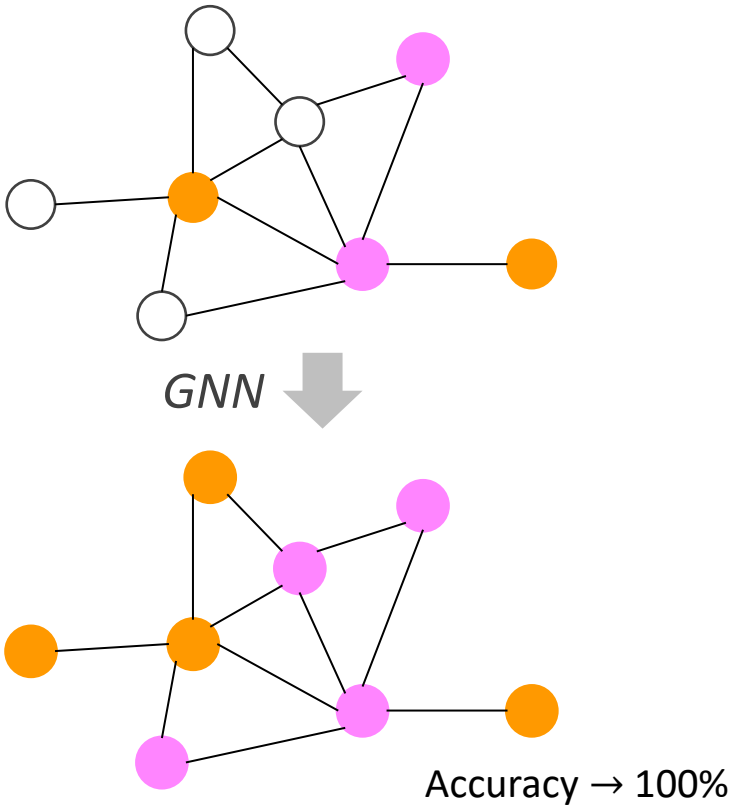
Target Label Classification (TLC)



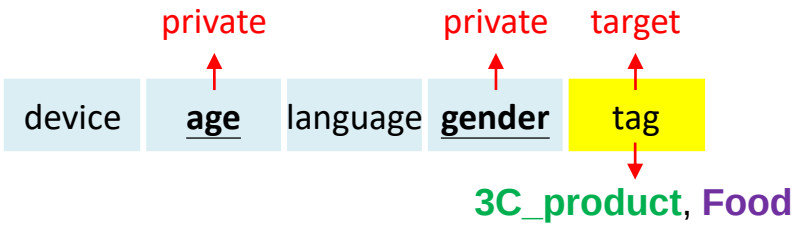


Privacy Attack =

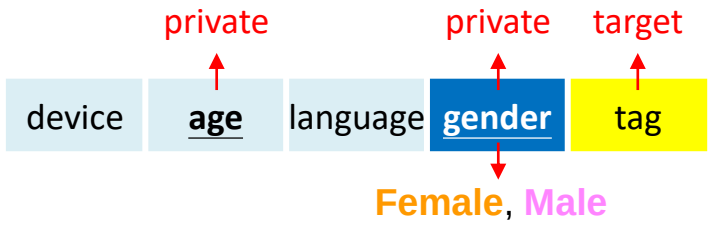
Private Label Classification (PLC)



Why can PLC work?



Target Label Classification (TLC)



Private Label Classification (PLC)

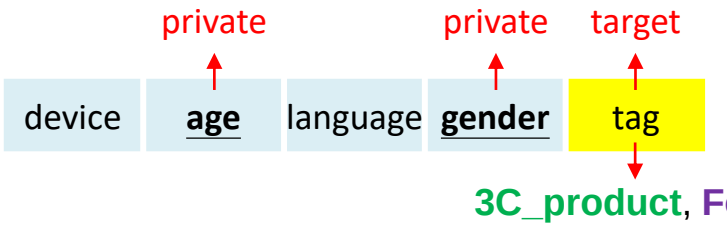
Not all of the users seriously care about their privacy!

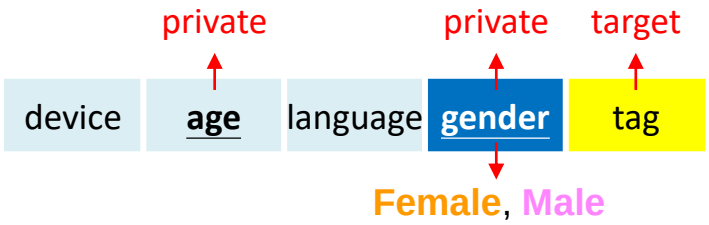
1) Some users may be not aware of privacy leaking
2) Some do not care whether private info is obtained by others, but are eager to promote themselves by showing full data

The adversary has some potential to collect sensitive data and further trains a PLC model

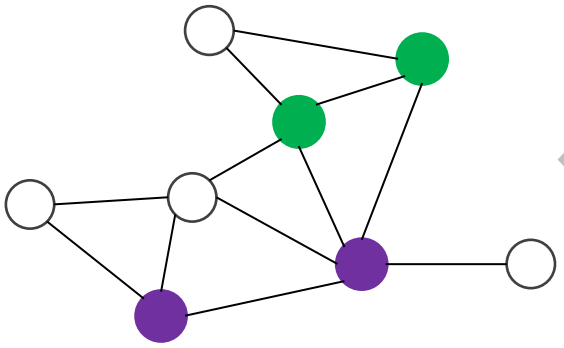
Our Goal: make PLC fail & keep TLC effective

Adversarial Defense against Privacy Attack

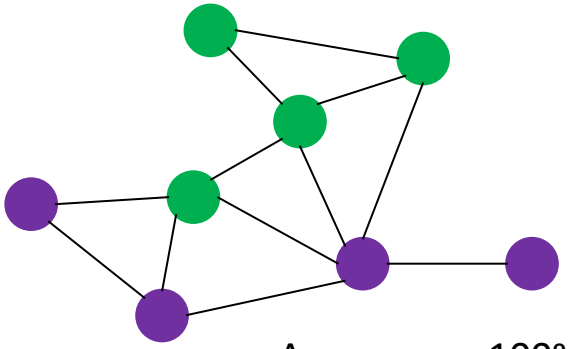




Target Label Classification on G'

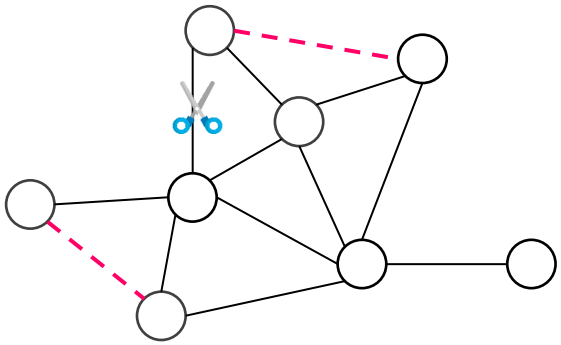


GNN

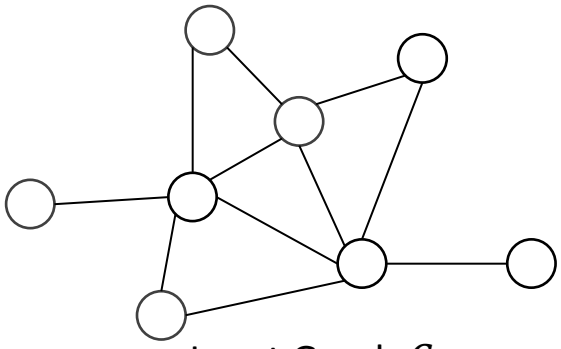


Accuracy → 100%

Perturbed Graph G'

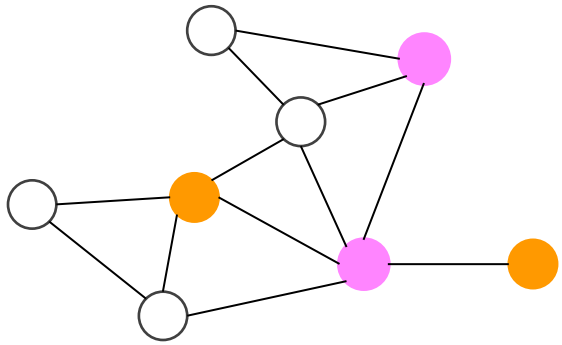


Unnoticeable Perturbation $G' \approx G$

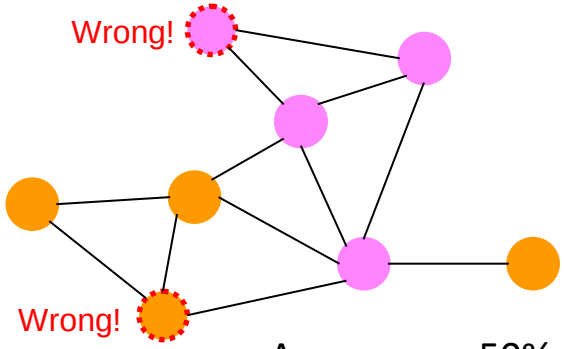


Input Graph G

Private Label Classification on G'



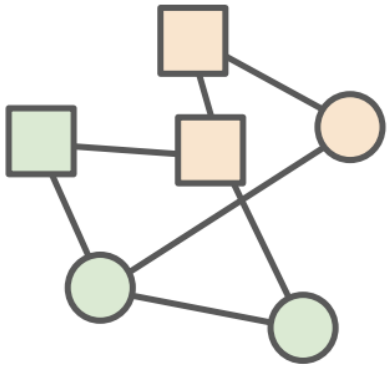
GNN



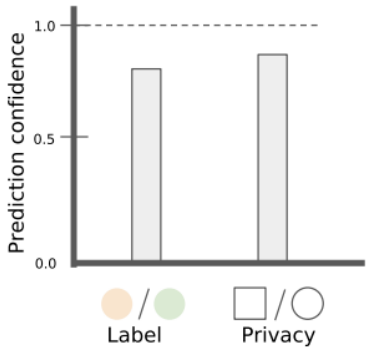
Accuracy → 50%

Privacy Defense on GNNs

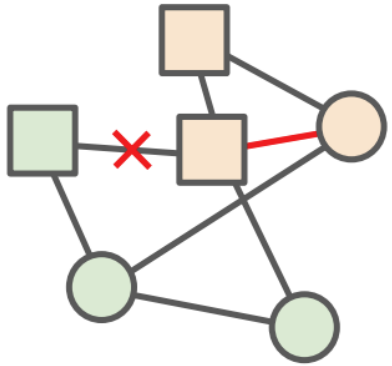
Input Graph

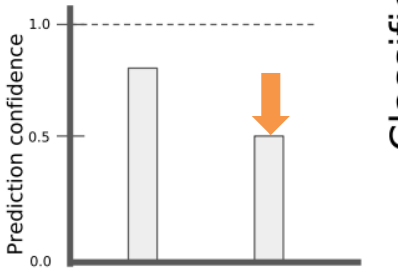


Node Classification Models



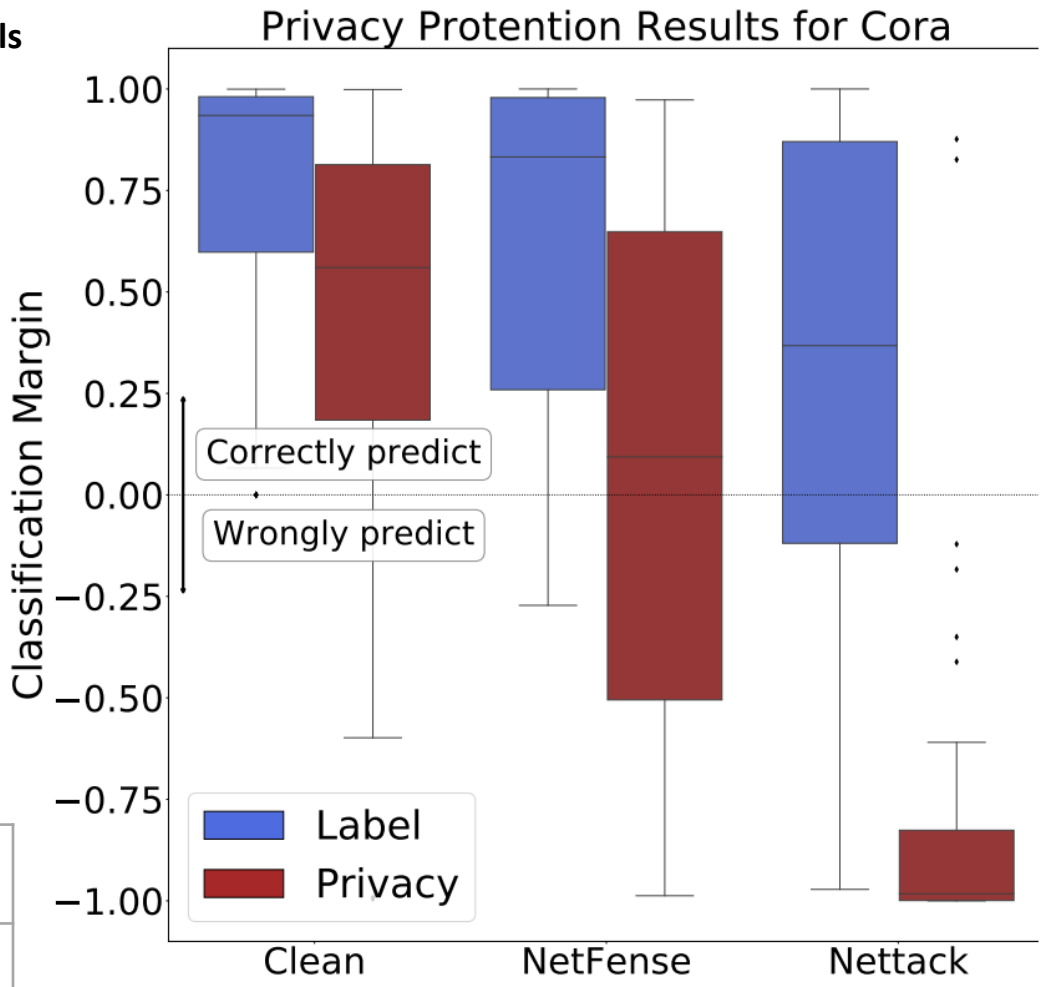
Privacy-preserving and unnoticeable perturbation





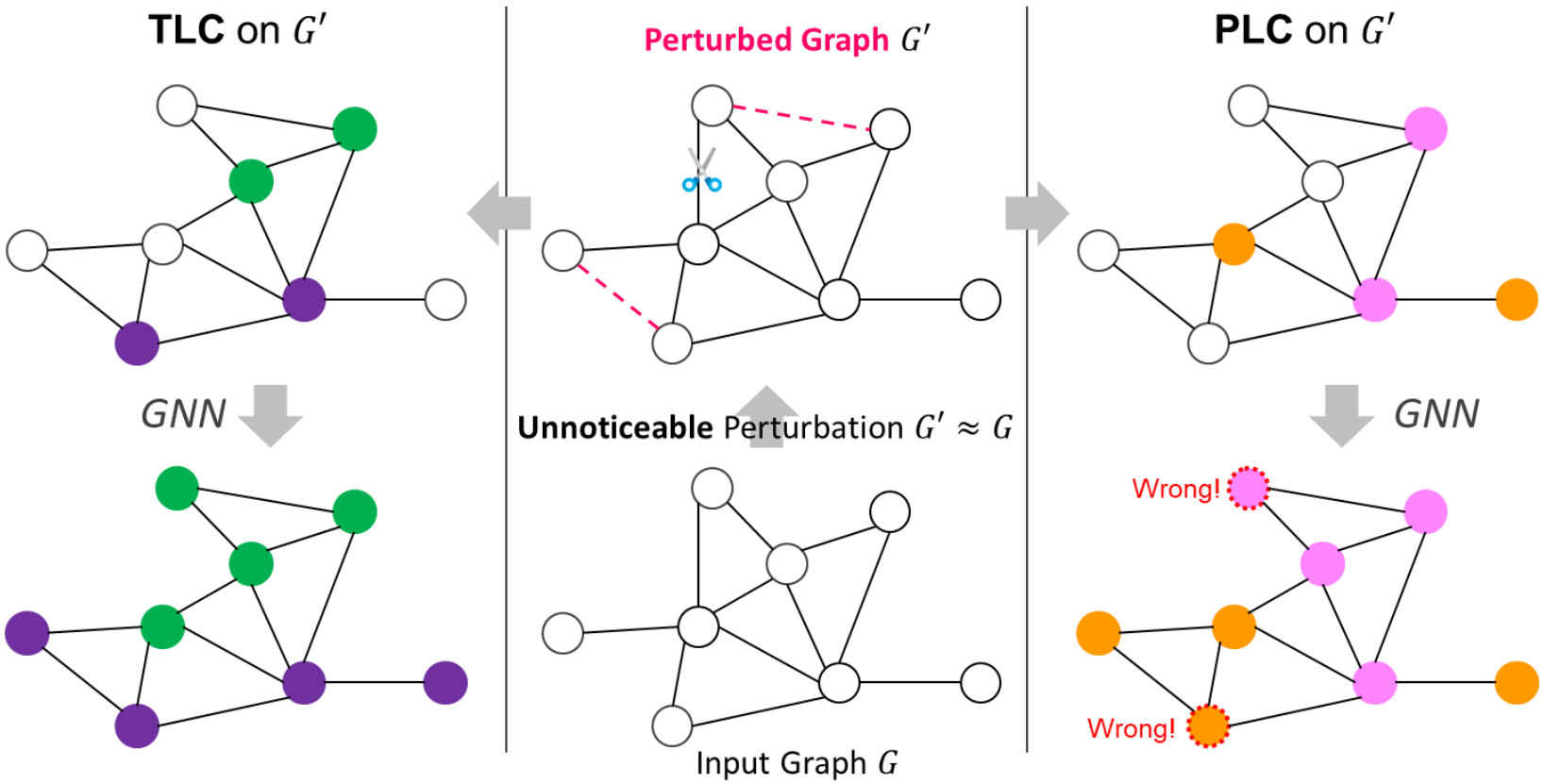
Label	Privacy
e.g., tags, interests	e.g., age, gender

Output: Perturbed Graph



Classification margin = Difference of prediction probs between ground-truth and 2-nd probable labels

Privacy Defense on GNNs



Perturb G to be G' such that $CM_{TLC}=|GNN_{TLC@t1}(G') - GNN_{TLC@t2}(G')|$ is **maximized**

Keep confident in TLC (effective in distinguishing target labels t1 and t2)

$CM_{PLC}=|GNN_{PLC@p1}(G') - GNN_{PLC@p2}(G')|$ is **minimized**

(i.e., close to 0)

Have minimum confident in PLC (fail to distinguish private labels p1 and p2)

Evaluation Settings

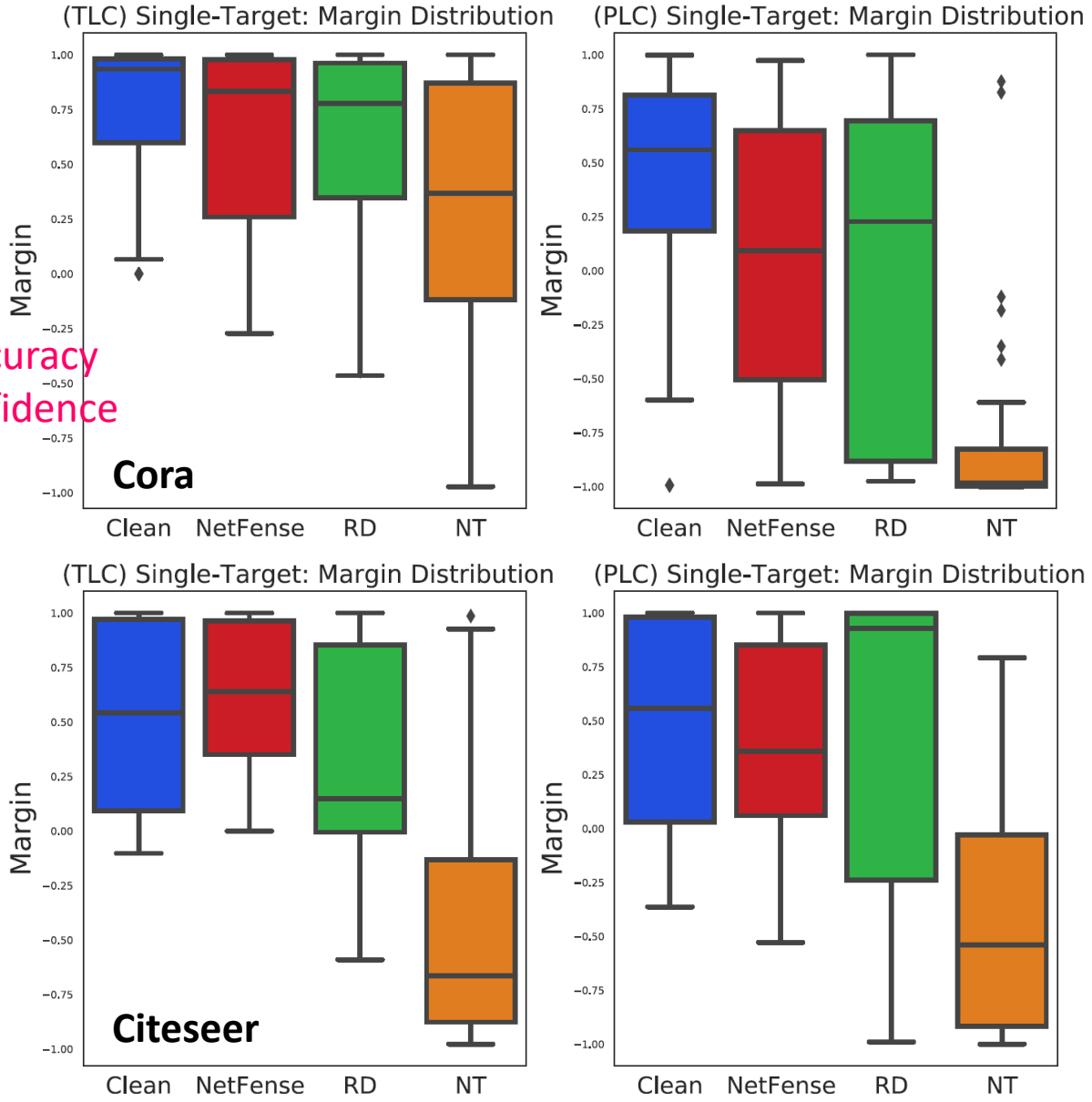
- 3 Typical graph datasets
- TLC: multiple-class classification
- PLC: binary classification
- Train/val/test = 0.1/0.1/0.8 (repeat 20 times)
- Settings of privacy attacks

Dataset	#nodes	#edges
Cora	2,708	5,429
Citeseer	3,312	4,715
PIT	851	16,392

	Single Target	**Multiple** Targets
Budget (# pairs of nodes)	**20** per target	**10** per target
Target Selection	10 nodes with highest/lowest CM, and random 20 nodes	All test nodes
Evaluation (avg CM)	Perturbed nodes	[Set] Perturbed nodes [Overall] All test nodes

Single-Target Results

Maintain TLC accuracy
Destroy PLC confidence



Robustness Issue on Fakes & Frauds

	Robustness Issue	Real-world Cases
Fake News	early detection, label noise & scarcity, edge noise& sparsity	1) Need fast & accurate detection before wide spread 2) Costly and labor-intensive of labeling 3) Human labeling may be incorrect (mis-led by bad guys) 4) Comments and re-shared can be manipulated
Customs Frauds	label scarcity, unseen users/items	1) Low inspection rate 2) Always have new/unseen importers and HS codes
Food Frauds	no labels, high-dim & noisy features	1) Costly and labor-intensive of labeling 2) Number of chemicals is huge 3) Chemical names are noisy and complicated
Link Frauds	adversarial privacy attacks	1) Bad guys can manipulate connections 2) Node classification can infer private attributes

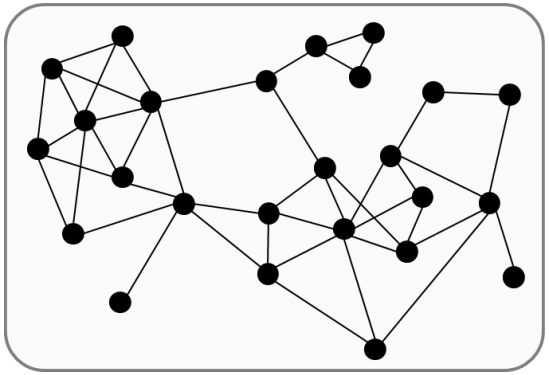
Warp Up

	Detection	Robust Detection via GraphML		
		Methods	Robustness Issue	GraphML
Fake News	GCAN [1] [ACL'20]	RERD [2] [TKDD'24]	early detection, label noise & scarcity, edge noise& sparsity	Multi-view Information Propagation GNNs
Customs Frauds	DATE [3] [KDD'20]	GraphFC [4] [CIKM'23]	label scarcity, unseen users/items	Self-supervised Heterogeneous GNNs
Food Frauds	Existing Supervised Methods	GraphCAR [5] [WWW'24]	no labels, high-dim & noisy features	Bipartite GNNs with Self-supervised Contrastive Learning
Link Frauds		NetFense [6] [TKDE'23]	adversarial privacy attacks	Competitive Augmentation & GNN Bi-optimization

1. Lu and Li. "GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media." ACL 2020.
2. Zhuang et al. "Towards Robust Rumor Detection with Graph Contrastive and Curriculum Learning." ACM TKDD 2024.
3. Kim and Tsai et al. "DATE: Dual Attentive Tree-aware Embedding for Customs Frauds Detection." ACM KDD 2020.
4. Tsai et al. "GraphFC: Customs Fraud Detection with Label Scarcity." ACM CIKM 2023.
5. Yang et al. "Detecting Illicit Food Factories from Chemical Declaration Data via Graph-aware Self-supervised Contrastive Anomaly Ranking." ACM TheWebConf (WWW) 2024.
6. Hsieh and Li. "NetFense: Adversarial Defenses against Privacy Attacks on Neural Networks for Graph Data." IEEE TKDE 2023.

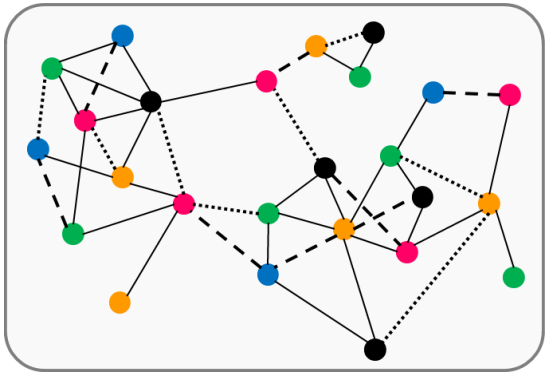
Key: Formulating the Graph

Homogeneous Graph



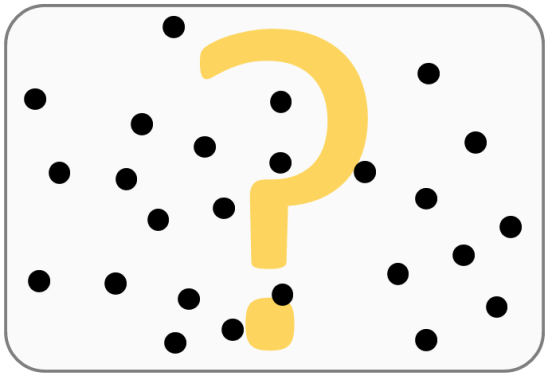
- Information propagation on social media
- Social networks
- Financial networks

Heterogeneous Graph



- User-item interactions (e.g., factory-chemical declarations)
- Transactions-metadata relations (e.g., declaration-importer-hscode)

Invisible Graph

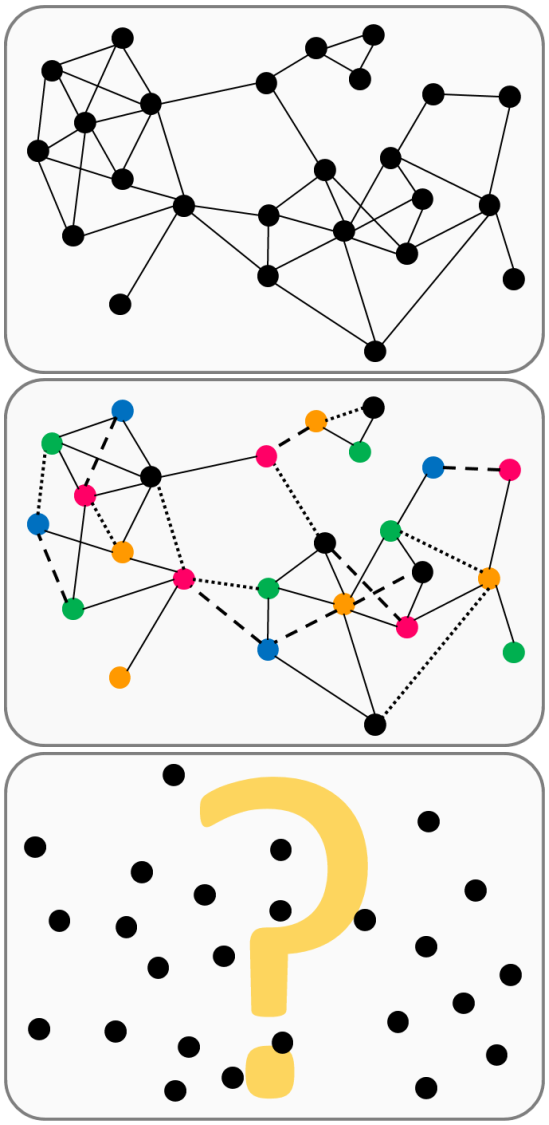


Construct edges by kNN or GSL

- Text graphs (e.g., news documents)
- Time-series graphs (e.g., sensors)
- Image graphs (e.g., deepfake images)

X: Graph Data

Key: Robust Graph Learning



f( )

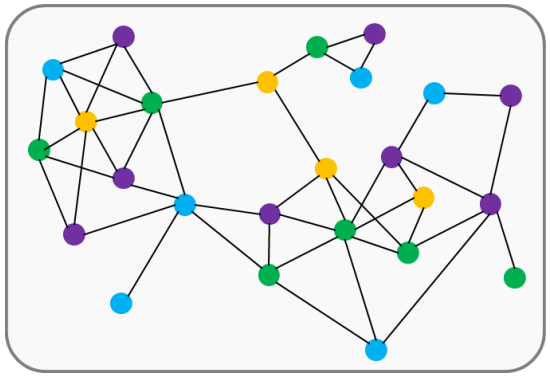
f=GNN

X: Graph Data

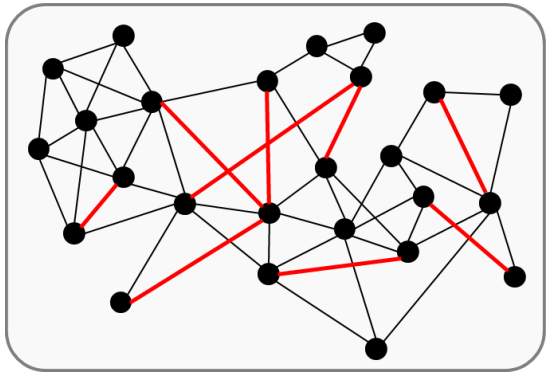
- Propagate the supervision signals
- Leverage label correlation
- Enhance label usefulness

Fine-tuning Task:
Node Classification

Y: Labels on nodes (color)



Pre-Training Task:
Link Prediction

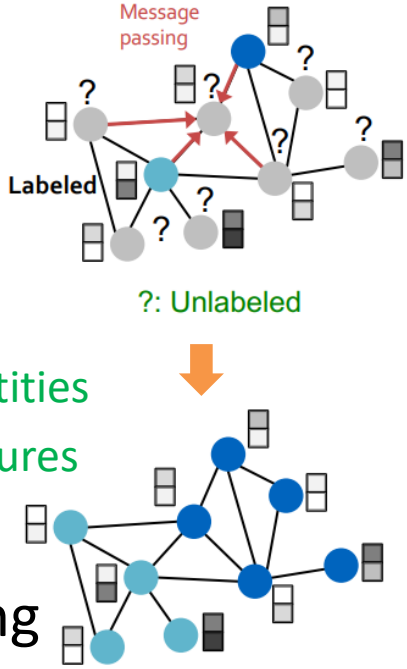


Y: Link existence (red lines)

- Spread useful features
- Correct noisy signals
- Capture feature correlation
- Warm up representation learning

Final Remarks

- **Why Useful is Robust Graph ML in combating fakes & frauds?**
  - Incorporate graph structure with node features
  - Connect the dots
    - Depict and exploit the relationships between instances
  - Message passing
    - Propagate supervised signals from labeled to unlabeled nodes
    - Propagate useful features from clean to noisy nodes
  - Self-correlation modeling
    - Reconstruct graph structure – capture correlation between entities
    - Reconstruct feature values – capture correlation between features
  - Graph augmentation – enhance message passing and correlation modeling
  - High-order interactions between entities
    - Better distinguish instances in a finer-grained manner



Can We Win the War on Combating Fakes and Frauds?

- Technically: need to improve the **robustness** of various aspects
- In fact, it's a complex problem, no easy solution
- Need "**collective**" intelligence
  - Social media, technology companies
  - Governments
  - International organizations
  - Civil society: journalists, fact-checkers, media, NGOs, etc.
  - Researchers: academia, industry

Bridging Trust: Combating Fakes and Frauds with Robust Graph Learning

Cheng-Te Li

Professor
National Cheng Kung Univ.

Email chengte@ncku.edu.tw

Web https://sites.google.com/view/chengteli/